THE ANNALS *of* MATHEMATICAL STATISTICS

(FOUNDED BY H. C. CARVER)
THE OFFICIAL JOURNAL OF THE INSTITUTE OF MATHEMATICAL STATISTICS

VOLUME XIV

1943

THE ANNALS OF MATHEMATICAL STATISTICS

QA
311
ANN

EDITED BY

S. S. WILKS, *Editor*

A. T. CRAIG	H. HOTELLING
W. E. DEMING	J. NEYMAN
T. C. FRY	W. A. SHEWHART

WITH THE COÖPERATION OF

W. G. COCHRAN	P. S. DWYER	P. G HOEL
J. H. CURTISS	C EISENHART	W. G. MADOW
H. F. DODGE	W. K. FELLER	A. WALD

The ANNALS OF MATHEMATICAL STATISTICS is published quarterly by the Institute of Mathematical Statistics Mt. Royal & Guilford Aves., Baltimore, Md. Subscriptions, renewals, orders for back numbers and other business communications should be sent to the ANNALS OF MATHEMATICAL STATISTICS, Mt. Royal & Guilford Aves., Baltimore, Md., or to the Secretary of the Institute of Mathematical Statistics, E. G. Olds, Carnegie Institute of Technology, Pittsburgh, Pa. Changes in mailing address which are to become effective for a given issue should be reported to the Secretary on or before the 15th of the month preceding the month of that issue. The months of issue are March, June, September and December. Because of war-time difficulties of publication, issues may often be from two to four weeks late in appearing. *Subscribers are therefore requested to wait at least 30 days after month of issue before making inquiries concerning non-delivery.*

Manuscripts for publication in the ANNALS OF MATHEMATICAL STATISTICS should be sent to S. S. Wilks, Fine Hall, Princeton, New Jersey. Manuscripts should be typewritten double-spaced with wide margins, and the original copy should be submitted. Footnotes should be reduced to a minimum and whenever possible replaced by a bibliography at the end of the paper; formulae in footnotes should be avoided. Figures, charts, and diagrams should be drawn on plain white paper or tracing cloth in black India ink twice the size they are to be printed. Authors are requested to keep in mind typographical difficulties of complicated mathematical formulae.

Authors will ordinarily receive only galley proofs. Fifty reprints without covers will be furnished free. Additional reprints and covers furnished at cost.

The subscription price for the ANNALS is $5.00 per year. Single copies $1.50. Back numbers are available at $5.00 per volume, or $1.50 per single issue.

COMPOSED AND PRINTED AT THE
WAVERLY PRESS, INC.
BALTIMORE, MD., U. S. A.

Entered as second-class matter at the Post Office at Baltimore, Maryland, under the Act of March 3, 1879

CONTENTS OF VOLUME XIV

Articles

Anderson, T. W., and Villars, D. S. Some Significance Tests for Normal Bivariate Distributions 141

Anderson, T. W. On Card Matching 426

Bartky, Walter. Multiple Sampling with Constant Probability 363

Chung, Kai Lai. Generalization of Poincaré's Formula in the Theory of Probability 63

Chung, Kai Lai. On Fundamental Systems of Probabilities of a Finite Number of Events 123

Chung, Kai Lai. Further Results on Probabilities of a Finite Number of Events 234

Cochran, W. G. The Comparison of Different Scales of Measurement for Experimental Results 205

Curtiss, J. H. On Transformations Used in the Analysis of Variance 107

Dodge, H. F. A Sampling Inspection Plan for Continuous Production 264

Eisenhart, C., and Swed, Freda S. Tables for Testing Randomness of Grouping in a Sequence of Alternatives 66

Feller, W. On a General Class of "Contagious" Distributions 389

Gumbel, E. J. On Serial Numbers 163

Gumbel, E. J. On the Reliability of the Classical Chi-Square Test 253

Hansen, Morris H., and Hurwitz, William N. On the Theory of Sampling from Finite Populations 333

Hoel, Paul. On Indices of Dispersion 155

Hoel, Paul. The Accuracy of Sampling Methods in Ecology 289

Hotelling, Harold. Some New Methods in Matrix Calculation 1

Hurwitz, William N., and Hansen, Morris H. On the Theory of Sampling from Finite Populations 333

Mann, H. B., and Wald, A. On Stochastic Limit and Order Relationships. 217

Mann, H. B. On the Construction of Sets of Orthogonal Latin Squares 401

Mathisen, Harold C. A Method of Testing the Hypothesis that Two Samples are from the Same Population 188

Mood, Alexander M. On the Dependence of Sampling Inspection Plans upon Population Distributions 415

Peiser, Alfred M. Asymptotic Formulas for Significance Levels of Certain Distributions 56

Samuelson, Paul A. Fitting General Gram-Charlier Series 179

Scheffé, Henry. On Solutions of the Behrens-Fisher Problem Based on the t-Distribution 35

Scheffé, Henry. On a Measure Problem Arising in the Theory of Non-Parametric Tests 227

Scheffé, Henry. Statistical Inference in the Non-Parametric Case 305

THE ANNALS OF MATHEMATICAL STATISTICS

QA
311
ANN

EDITED BY

S. S. WILKS, *Editor*

A. T. CRAIG	H. HOTELLING
W. E. DEMING	J. NEYMAN
T. C. FRY	W. A. SHEWHART

WITH THE COÖPERATION OF

W. G. COCHRAN	P. S. DWYER	P. G. HOEL
J. H. CURTISS	C EISENHART	W. G. MADOW
H. F. DODGE	W. K. FELLER	A. WALD

The ANNALS OF MATHEMATICAL STATISTICS is published quarterly by the Institute of Mathematical Statistics Mt. Royal & Guilford Aves., Baltimore, Md. Subscriptions, renewals, orders for back numbers and other business communications should be sent to the ANNALS OF MATHEMATICAL STATISTICS, Mt. Royal & Guilford Aves., Baltimore, Md., or to the Secretary of the Institute of Mathematical Statistics, E. G. Olds, Carnegie Institute of Technology, Pittsburgh, Pa. Changes in mailing address which are to become effective for a given issue should be reported to the Secretary on or before the 15th of the month preceding the month of that issue. The months of issue are March, June, September and December. Because of war-time difficulties of publication, issues may often be from two to four weeks late in appearing. *Subscribers are therefore requested to wait at least 30 days after month of issue before making inquiries concerning non-delivery.*

Manuscripts for publication in the ANNALS OF MATHEMATICAL STATISTICS should be sent to S. S. Wilks, Fine Hall, Princeton, New Jersey. Manuscripts should be typewritten double-spaced with wide margins, and the original copy should be submitted. Footnotes should be reduced to a minimum and whenever possible replaced by a bibliography at the end of the paper; formulae in footnotes should be avoided. Figures, charts, and diagrams should be drawn on plain white paper or tracing cloth in black India ink twice the size they are to be printed. Authors are requested to keep in mind typographical difficulties of complicated mathematical formulae.

Authors will ordinarily receive only galley proofs. Fifty reprints without covers will be furnished free. Additional reprints and covers furnished at cost.

The subscription price for the ANNALS is $5.00 per year. Single copies $1.50. Back numbers are available at $5.00 per volume, or $1.50 per single issue.

COMPOSED AND PRINTED AT THE
WAVERLY PRESS, INC.
BALTIMORE, MD., U. S. A.

Entered as second-class matter at the Post Office at Baltimore, Maryland, under the Act of March 3, 1879

CONTENTS OF VOLUME XIV

Articles

Anderson, T. W., and Villars, D. S. Some Significance Tests for Normal Bivariate Distributions 141

Anderson, T. W. On Card Matching 426

Bartky, Walter. Multiple Sampling with Constant Probability 363

Chung, Kai Lai. Generalization of Poincaré's Formula in the Theory of Probability 63

Chung, Kai Lai. On Fundamental Systems of Probabilities of a Finite Number of Events 123

Chung, Kai Lai. Further Results on Probabilities of a Finite Number of Events 234

Cochran, W. G. The Comparison of Different Scales of Measurement for Experimental Results 205

Curtiss, J. H. On Transformations Used in the Analysis of Variance 107

Dodge, H. F. A Sampling Inspection Plan for Continuous Production 264

Eisenhart, C., and Swed, Freda S. Tables for Testing Randomness of Grouping in a Sequence of Alternatives 66

Feller, W. On a General Class of "Contagious" Distributions 389

Gumbel, E. J. On Serial Numbers 163

Gumbel, E. J. On the Reliability of the Classical Chi-Square Test 253

Hansen, Morris H., and Hurwitz, William N. On the Theory of Sampling from Finite Populations 333

Hoel, Paul. On Indices of Dispersion 155

Hoel, Paul The Accuracy of Sampling Methods in Ecology 289

Hotelling, Harold. Some New Methods in Matrix Calculation 1

Hurwitz, William N., and Hansen, Morris H. On the Theory of Sampling from Finite Populations 333

Mann, H. B., and Wald, A. On Stochastic Limit and Order Relationships 217

Mann, H. B. On the Construction of Sets of Orthogonal Latin Squares 401

Mathisen, Harold C. A Method of Testing the Hypothesis that Two Samples are from the Same Population 188

Mood, Alexander M. On the Dependence of Sampling Inspection Plans upon Population Distributions 415

Peiser, Alfred M. Asymptotic Formulas for Significance Levels of Certain Distributions 56

Samuelson, Paul A. Fitting General Gram-Charlier Series 179

Scheffé, Henry. On Solutions of the Behrens-Fisher Problem Based on the t-Distribution 35

Scheffé, Henry. On a Measure Problem Arising in the Theory of Non-Parametric Tests 227

Scheffé, Henry. Statistical Inference in the Non-Parametric Case 305

SIMON, HERBERT A. Symmetric Tests of the Hypothesis that the Mean of One Normal Population Exceeds that of Another..... 149
SWED, FREDA S., AND EISENHART, C. Tables for Testing Randomness of Grouping in a Sequence of Alternatives..... 66
VILLARS, D S., AND ANDERSON, T. W. Some Significance Tests for Normal Bivariate Distributions..... 141
VON MISES, R. On the Problem of Testing Hypotheses..... 238
WALD, ABRAHAM. An Extension of Wilks' Method of Setting Tolerance Limits..... 45
WALD, ABRAHAM. On the Efficient Design of Statistical Investigations... 134
WALD, A., AND MANN, H. B. On Stochastic Limit and Order Relationships. 217
WALD, A., AND WOLFOWITZ, J. An Exact Test for Randomness in the Non-Parametric Case Based on Serial Correlation..... 378
WOLFOWITZ, J. On the Theory of Runs with Some Applications to Quality Control..... 280
WOLFOWITZ, J., AND WALD, A. An Exact Test for Randomness in the Non-Parametric Case Based on Serial Correlation..... 378

NOTES

AROIAN, LEO A. A New Approximation to the Levels of Significance of the Chi-Square Distribution..... 93
CRAIG, A. T. A Note on the Best Linear Estimate..... 88
CRAIG, A. T. Note on the Independence of Certain Quadratic Forms..... 195
DORFMAN, ROBERT. The Detection of Defective Members of Large Populations..... 436
HOTELLING, HAROLD. Further Points on Matrix Calculation and Simultaneous Equations..... 440
PAULSON, EDWARD. A Note on Tolerance Limits..... 90
KAPLANSKY, IRVING. A Characterization of the Normal Distribution.... 197

MISCELLANEOUS

Abstracts of Papers..... 445
Annual Report of the President of the Institute..... 98
Annual Report of the Secretary-Treasurer of the Institute..... 100
Constitution and By-Laws of the Institute..... 102
Directory of the Institute..... 448
News and Notices..... 96, 199, 301, 442
Report on the New Brunswick Meeting of the Institute..... 444
Report on the New York Meeting of the Institute..... 204
Report on the First Meeting of the Pittsburgh Chapter of the Institute.... 303
Report on the Second Meeting of the Pittsburgh Chapter of the Institute.. 445
Report on the Washington Meeting of the Institute..... 302
Special Courses in Statistical Quality Control..... 202

THE ANNALS *of* MATHEMATICAL STATISTICS

(FOUNDED BY H. C. CARVER)

THE OFFICIAL JOURNAL OF THE INSTITUTE OF MATHEMATICAL STATISTICS

Contents

Some New Methods in Matrix Calculation. HAROLD HOTELLING 1

On Solutions of the Behrens-Fisher Problem, Based on the t-Distribution. HENRY SCHEFFÉ 35

An Extension of Wilk's Method of Setting Tolerance Limits. ABRAHAM WALD 45

Asymptotic Formulas for Significance Levels of Certain Distributions. ALFRED M. PEISER 56

Generalization of Poincaré's Formula in the Theory of Probability. KAI LAI CHUNG 63

Tables for Testing Randomness of Grouping in a Sequence of Alternatives. FRIEDA S. SWED AND C. EISENHART 66

Notes:

A Note on the Best Linear Estimate. A. T. CRAIG 88

A Note on Tolerance Limits. EDWARD PAULSON 90

A New Approximation to the Levels of Significance of the Chi-Square Distribution. LEO A. AROIAN 93

News and Notices 96

Annual Report of the President of the Institute 98

Annual Report of the Secretary-Treasurer of the Institute 100

Constitution and By-Laws of the Institute 102

Vol. XIV, No. 1 — March, 1943

THE ANNALS OF MATHEMATICAL STATISTICS

EDITED BY

S. S. WILKS, *Editor*

A. T. CRAIG
W. E. DEMING
T. C. FRY
H. HOTELLING
J. NEYMAN
W. A. SHEWHART

WITH THE COÖPERATION OF

W. G. COCHRAN
J. H. CURTISS
H. F. DODGE
P. S. DWYER
C. EISENHART
W. K. FELLER
P. G. HOEL
W. G. MADOW
A. WALD

The ANNALS OF MATHEMATICAL STATISTICS is published quarterly by the Institute of Mathematical Statistics, Mt. Royal & Guilford Aves., Baltimore, Md. Subscriptions, renewals, orders for back numbers and other business communications should be sent to the ANNALS OF MATHEMATICAL STATISTICS, Mt. Royal & Guilford Aves., Baltimore, Md., or to the Secretary of the Institute of Mathematical Statistics, E. G. Olds, Carnegie Institute of Technology, Pittsburgh, Pa. Changes in mailing address which are to become effective for a given issue should be reported to the Secretary on or before the 15th of the month preceding the month of that issue. The months of issue are March, June, September and December. Because of war-time difficulties of publication, issues may often be from two to four weeks late in appearing. *Subscribers are therefore requested to wait at least 30 days after month of issue before making inquiries concerning non-delivery.*

Manuscripts for publication in the ANNALS OF MATHEMATICAL STATISTICS should be sent to S. S. Wilks, Fine Hall, Princeton, New Jersey. Manuscripts should be typewritten double-spaced with wide margins, and the original copy should be submitted. Footnotes should be reduced to a minimum and whenever possible replaced by a bibliography at the end of the paper; formulae in footnotes should be avoided. Figures, charts, and diagrams should be drawn on plain white paper or tracing cloth in black India ink twice the size they are to be printed. Authors are requested to keep in mind typographical difficulties of complicated mathematical formulae.

Authors will ordinarily receive only galley proofs. Fifty reprints without covers will be furnished free. Additional reprints and covers furnished at cost.

The subscription price for the ANNALS is $5.00 per year. Single copies $1.50. Back numbers are available at $5.00 per volume, or $1.50 per single issue.

COMPOSED AND PRINTED AT THE
WAVERLY PRESS, INC.
BALTIMORE, MD., U. S. A.

Entered as second-class matter at the Post Office at Baltimore, Maryland, under the Act of March 3, 1879

By action of the Board of Directors of the Institute, The 1943 Volume of the Annals of Mathematical Statistics is respectfully dedicated to

HENRY LEWIS REITZ

on the occasion of his retirement after twenty-five years of service as Head of the Department of Mathematics of the University of Iowa, and in recognition of his contributions to the initiation and development of mathematical statistics in America.



M. L. Rieß

SOME NEW METHODS IN MATRIX CALCULATION[1]

By Harold Hotelling

Columbia University

I. Introduction

1. The increased practical importance of matrix calculation. This paper will be concerned chiefly with methods of finding the inverse of a matrix, and of finding the latent roots and latent vectors, which are also known by a variety of other names associated with particular applications, such as principal axes in geometry and mechanics, and principal components in psychology. These two computational problems are of extremely wide application. The first is closely related to the solution of systems of linear equations, which we shall also consider. In the method of least squares the solution of the normal equations is best carried out with the help of the inverse of the matrix of the coefficients, since at least some of the elements of this inverse matrix are needed in evaluating the results in terms of probability, a vitally necessary step, and since the inverse matrix is useful also in various other ways, such as altering the set of predictors used in a regression equation. Modern statistics also utilizes quadratic and bilinear forms such as the generalized Student ratio [15] for discriminating between samples according to multiple variates instead of one only, the associated discriminant functions [10], the closely related figurative distance of Mahalanobis, Bose and Roy [5] and the critical statistic in an investigation by Wald [28] of the efficient classification of an individual into one of two groups. All these may be calculated very easily from the inverse of a matrix of sums of products, or of covariances or correlations, or from the principal components. Consideration of the relations between two sets of variates [18] may utilize both the inverse of a matrix and a process resembling the calculation of principal components. Similar computational problems arise in applying to sets of numerous variates the contributions to multivariate statistical analysis of R. A. Fisher, S. S. Wilks, W. G. Madow, M. A Girshick, P. L. Hsu and M. S. Bartlett. Among the non-statistical applications of the inverse matrix and of latent roots and vectors are problems of dynamics, both in astronomy and in airplane design [12], the analysis of stresses and strains in structures [26, 27], and electrical engineering problems [24].

Perhaps no objection to attempts at statistical inference is more common than that the variation of this or that relevant factor has been ignored. For example in dealing with time series the need of allowing for trend and seasonal variation, perhaps by means of a sequence of orthogonal polynomials for trend and of

[1] Revision of a paper presented at the Symposium on Numerical Calculation held Dec. 28, 1941 in New York by the Institute of Mathematical Statistics and the American Statistical Association with the cooperation of the Committee on Addresses in Applied Mathematics of the American Mathematical Society. For the program of the Symposium see the *Annals of Mathematical Statistics* for March, 1942, p. 103.

trigonometric functions for seasonal variation, is well recognized. It is indeed desirable to use regression equations with a liberal number of predictors to eliminate spurious influences, as well as to reduce the error variance, and likewise in other statistical methods. But the computational difficulties in the joint analysis of the desired number of variables have frequently seemed too formidable. We shall see how efficient techniques, in conjunction with efficient machines, can go far to facilitate the use of an appropriate number of variables by reducing the labor to modest dimensions.

While the rise of modern multivariate statistical theory has made available new exact tests of hypotheses in terms of probability over a wide range of cases in which multiple measurements are involved, such measurements have been accumulating on a large scale. In many psychological, anthropometric, astronomical, meteorological and economic fields, actual measurements are available on numbers of variates far greater than have been regarded as amenable, within practical limits, to adequate treatment by the numerical methods generally used. In some instances the number of cases in which complete sets of these variates are available is also large. The 1931 census of India included an extensive sample in which fifty physical variates were measured for each individual. Karl J. Holzinger and his collaborators have worked out and circulated privately a complete matrix of correlations among 78 mental tests. Astronomers have indicated the desirability of a recalculation of the elements of the solar system by means of a gigantic least-square solution with 150 or more unknowns, at the same time deploring the seeming impossibility of this ever being carried out. To apply the methods of modern theoretical statistics to derive from such observations all the important information they contain is an enterprise whose feasibility depends on new numerical methods.

The chief computational problems, apart from those of tabulating and providing convenient approximations for the probability distributions, are (1) the calculation of the many sums of products of pairs of p variates when p is large, and (2) operations on the matrices of these sums of products such as finding the inverse and the principal components. The first problem, which in classical applications of the method of least squares to long series has seemed the heavier, has in a sense been solved by the use of punched cards. A card is used for each case, and all p variates are punched into it. By running the cards repeatedly through a machine wired at each run to select a particular pair of variates, multiply them together, and cumulate the products, this part of the work may be disposed of with great speed. The cost of the machines does at present limit the economical use of this method to rather large numbers, both of variates and of cases. This limit has recently been pushed upward by the introduction of improved multiplying calculators, with high-speed automatic multiplication and squaring locks. But these mechanical advances, in combination with recent discoveries in statistical theory, the increasingly felt need to resort to numerous variates, and the actual existence in many cases of data on such multiple variates, emphasize the need for rapid, economical and accurate calculations with matrices whose elements are sums of products.

Modern machine methods, especially those of the punched-card type, but also those using machines such as the Monroe, Marchant and Fridén, tend to reduce the work of formation of sums of products, in comparison with other operations, to such an extent as to enhance the relative value of methods in which such calculation of direct product-sums is important. Thus products of matrices are much simpler to compute than inverses, and positive than negative powers. Indeed, powers and products of matrices can be computed by means of punched-card machines, and for large matrices this is doubtless the most efficient procedure now available, though considerable rewiring is needed. There is also a possibility, which does not seem too remote, of development of further devices to do this rewiring automatically.

2. Iterative and direct methods. Partitioned matrices. In later sections we shall deal chiefly with certain iterative methods, giving particular attention to the neglected question of limits of error in stopping at any point, and considering the rate of approach to the desired solution. For finding the roots of a matrix and the associated vectors, if the matrix has more than about four rows, it seems clear that an iterative method is the most economical of labor in all but very special cases. On the other hand the problems of solving systems of linear equations and finding the inverse of a matrix do not usually yield readily to iterative methods unless an approximation to the solution is available to begin with. This approximation is not necessarily a very close one, but must not be too wild. It may in some cases be obtained from a general knowledge of the subject.

The Mallock electrical device [22] is capable of solving almost instantaneously ten linear equations in ten unknowns with perhaps two significant digits in each result, though this question of accuracy remains to be elucidated. The combination of this device with the iterative method of Section 7 below, and with the use of partitioning for matrices of more than ten rows, offers what seems at present the best hope for the systematic inversion of large matrices. Since only one of the Mallock machines is in existence (it is in Cambridge, England), some adaptation of the Doolittle or related methods will ordinarily be used. By taking advantage of the possibilities in modern calculating machines of accumulating products to reduce the amount of writing required in the Doolittle method, exceedingly compact and efficient methods have been developed for solving systems of linear equations and for evaluating inverse matrices by Dwyer [7, 8, 9], who utilized the earlier work of Waugh, Kurtz, Horst, Dunlap and Cureton cited by him, and for solving systems of linear equations, by Crout [6]. Dwyer gives valuable bibliographies.

By some of these methods, or from a general knowledge of the subject, one may well obtain approximate solutions correct to a very small number of decimal places, and then by iteration get as many more places as are required, with labor far less than would be necessary to carry through from the beginning the requisite number of places. Further applications of iterative methods arise when a least-square solution is to be revised, either on account of new observa-

tions or because of errors discovered in the original observations or calculations. But however a least-square calculation or the evaluation of any inverse matrix begins, and whatever intermediate steps are taken, it seems advisable to terminate it with the method of Section 7. This combines a check on the previous work, at a labor cost equivalent merely to substituting the values found for the unknowns into the equations, with an improvement in accuracy and a useful limit of error for the unknowns.

In the inversion of large matrices there are important possibilities in the properties of partitioning. For example, a square matrix of $2p$ rows may be partioned into four square matrices a, b, c, d, of p rows, and written

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

If this is multiplied on the right by another partitioned square matrix of $2p$ rows which may be written

$$\begin{bmatrix} A & C \\ B & D \end{bmatrix},$$

where A, B, C, D are square p-rowed matrices, the product

$$\begin{bmatrix} aA + bB & aC + bD \\ cA + dB & cC + dD \end{bmatrix}$$

is identical with the result of partitioning the product of the two original $2p$-rowed matrices. If the second is the inverse of the first, this product is the identical matrix. Consequently, if the first matrix is given, we have for determining its inverse the four matrix equations in A, B, C, D,

$$aA + bB = 1 \qquad aC + bD = 0$$

$$cA + dB = 0 \qquad cC + dD = 1,$$

where 1 stands for the identical matrix of p rows and 0 for the p-rowed matrix consisting entirely of zeros. These equations may be solved just as in elementary algebra except that care must be used to perform matrix multiplications in correct order. Thus

$$A = (a - bd^{-1}c)^{-1}, \qquad B = -d^{-1}cA$$

$$D = (d - ca^{-1}b)^{-1} \qquad C = -a^{-1}bD.$$

These formulae call for inversion of four p-rowed matrices, namely d, $a - bd^{-1}c$, a, and $d - ca^{-1}b$. Without changing the number of such inversions we may choose alternative sets of matrices to invert, with economy of labor in certain cases. For example, if b is easy to invert, we may use for D the expression

$$D = b^{-1}aAbd^{-1}.$$

The formulae and numerical work are further simplified if the given matrix is symmetric. Other modes of partitioning are also possible, and may be valuable in various kinds of numerical work. Another method of obtaining the inverse of a matrix by partitioning is given by Frazer, Duncan and Collar [12, pp. 112-118], who also give an account of general properties of partitioned matrices. In the treatment of relations between two or more sets of variates [18, 31], partitioned matrices appear.

The most efficient method of calculation of a function of a matrix will depend in part on what else is to be calculated. For example, if the latent roots and vectors are needed for any reason as well as the inverse of a matrix, it is better to calculate the former first, and then the determination of the inverse matrix becomes a trivial task; but if the latent roots and vectors are not needed for some other purpose it is usually better not to calculate them but to use a more direct method to obtain the inverse. If in addition to the inverse the determinant is wanted, or many consecutive powers of a matrix, or if a matrix-multiplying machine considerably speedier than present procedures becomes available, a method [3] based on the Cayley-Hamilton theorem that a matrix satisfies its own characteristic equation may be recommended.

Iterative methods have what Whittaker and Robinson [30] call the pleasing characteristic that mistakes do not necessarily spoil the whole calculation, but tend to be corrected at later stages. This of course does not mean that there is no penalty for mistakes. They have an obvious tendency to prolong the number of repetitions required, and if repeated at late stages may actually prevent realization of a substantially correct result. A less obvious consequence of mistakes near the termination of an iterative calculation is that they tend to vitiate any limits of error that may be derived, including those that will be found below. Great care should be used to insure accurate calculation especially in the last stages of any iterative process.

To insure accuracy even before the last stages, and therefore efficiency, a check column consisting of the sums of the elements in the rows of matrices multiplied and added together may well be carried along. In multiplying two matrices only the check column of the second factor is used; it is multiplied by each row of the first factor to obtain the check column for the product. A computer thoroughly experienced with matrix multiplication may dispense with the check column at all stages but the last of an iterative process, relying on the self-correcting property of the process.

A simple but extremely valuable bit of equipment in matrix multiplication consists of two plain cards, with a re-entrant right angle cut out of one or both of them if symmetric matrices are to be multiplied. In getting the element of the ith row and jth column of the product, the ith row of the first factor and the jth column of the second should be marked by a card beside, above, or below it. In writing a symmetric matrix it is convenient to omit the elements below the principal diagonal. The re-entrant right angle is then utilized to mark off the numbers belonging to a particular row.

A report [13] on certain iterative methods of solving linear and other equations and of calculating latent roots and vectors, with engineering applications, was published by R. von Mises and H Geiringer in 1929 As part of a discussion of certain problems in psychology [16] the present author in 1933 described iterative processes both for solving systems of linear equations and for finding principal components, and later [17] showed how to accelerate convergence to principal components by repeatedly squaring the matrix. Further acceleration of convergence by other devices has been discovered by A. C. Aitken [2]. Dr. Geiringer has also discussed a method of solution of equations involving iteration by small groups of unknowns [14] The method of Kelley and Salisbury [20] should be noted. It has been used extensively by psychologists. Definite limits of error and measures of rate of convergence for this method are missing. Certain other iterative methods will be discussed in later sections. It will appear that the most-used methods are by no means the best.

Questions regarding the probability of a matrix of covariances satisfying particular conditions of computational significance may in some cases be illuminated with the help of the theory of the variates as a random sample of a larger aggregate. This theory was outlined in the latter part of the paper [16].

II. Linear Equations and Inverse Matrices

3. Accuracy of direct solution of linear equations. The question how many decimal places should be retained in the various stages of a least-square solution and of other calculations involving linear equations has been a puzzling one. It has not generally been realized how rapidly errors resulting from rounding may accumulate in the successive steps of such procedures as, for example, the Doolittle method. In this popular algorism for solving a system of equations

$$\sum_{j=1}^{p} a_{ij} x_j = g_i \qquad (i = 1, \cdots, p),$$

the equivalent of successive eliminations of $x_1, x_2, \cdots, x_{p-1}$ to obtain an equation in x_p alone is accomplished by calculating successively

$$a_{ij\cdot 1} = a_{ij} - a_{i1}a_{1j}/a_{11}, \qquad g_{i\cdot 1} = g_i - a_{i1}g_1/a_{11} \qquad (i, j = 2, 3, \cdots, p),$$

then

$$a_{ij\cdot 12} = a_{ij\cdot 1} - a_{i2\cdot 1}a_{2j\cdot 1}/a_{22\cdot 1},$$

$$g_{i\cdot 12} = g_{i\cdot 1} - a_{i2\cdot 1}g_{2\cdot 1}/a_{22\cdot 1} \qquad (i, j = 3, \cdots, p),$$

and so forth. Let us suppose that each of the a_{ij}'s and g_i's is subject to an error concerning which it is known only that its absolute value does not exceed ϵ. Thus if they are given accurately to k decimal places only, we have $\epsilon = 10^{-k}/2$. Let the actual errors be represented by δa_{ij} and δg_i. If these are small an estimate of the error in $g_{i\cdot 1}$ may be obtained by expanding in a Taylor series and retaining only the linear terms:

$$\delta g_{i\cdot 1} = \delta g_i - \frac{a_{i1}}{a_{11}}\,\delta g_1 - \frac{g_1}{a_{11}}\,\delta a_{i1} + \frac{a_{i1}g_1}{a_{11}^2}\,\delta a_{11}.$$

The closest upper bound for this error obtainable without special assumptions regarding the values of the given quantities is specified by the inequality

$$|\,\delta g_{i\cdot 1}\,| \leq \epsilon\left(1 + \left|\frac{a_{i1}}{a_{11}}\right| + \left|\frac{g_1}{a_{11}}\right| + \left|\frac{a_{i1}g_1}{a_{11}^2}\right|\right).$$

The a's and g's are often correlation coefficients. Any set of normal equations of least squares may be reduced to a form in which this is the case, and this reduction has considerable merits. The various correlation coefficients are frequently of interest in themselves, and their use in the normal equations practically insures that all the quantities appearing at any stage are of the same order of magnitude. This last is a very substantial advantage, partly because of the check column which is customarily carried along, in which each entry is the sum of the other entries in its row. Since the absolute value of a correlation coefficient is less than unity, and since a_{ii} becomes equal to unity, the last inequality gives in this case

$$|\,\delta g_{i\cdot 1}\,| < 4\epsilon,$$

and no closer inequality appears possible. In the same way we find for this case in which the a's are correlation coefficients that

$$|\,\delta a_{ij\cdot 1}\,| < 4\epsilon.$$

Proceeding from these inequalities in the same way, and neglecting the fact that $|\,a_{22\cdot 1}\,| < 1$ though like a_{11} it is put equal to unity in the argument, we find for the errors in $a_{ij\cdot 12}$ and $g_{i\cdot 12}$ the estimated upper bound 16ϵ, with an actual upper bound somewhat higher unless $a_{12} = 0$. Continuing in the same way we find for $a_{ij\cdot 12\cdots(p-1)}$ and $g_{i\cdot 12\cdots(p-1)}$ the estimated limit of error $4^{p-1}\epsilon$, with a possibility of a somewhat higher value up to $4^{p-1}\epsilon/a$, where a is the determinant $|\,a_{ij}\,| < 1$. The rapidity with which this increases with p is a caution against relying on the results of the Doolittle method or other similar elimination methods with any moderate number of decimal places when the number of equations and unknowns is at all large. Thus if $p = 11$ the limit of error exceeds a million times ϵ, indicating that if only one decimal place is wanted in the value of x_p the original correlations must be utilized to at least seven decimals, even if we neglect the additional errors introduced by dropping decimals beyond those retained in the intermediate stages of the calculation. The errors accumulate further during the back solution, so that if all the unknowns are wanted with one-place accuracy it is necessary to use the original correlations with substantially more than seven decimal places. For larger values of p the increase in the error limit is startling. Thus for $p = 27$ (the number of tests reported to be involved in a certain current procedure in classifying military personnel) the limit of error even for the first unknown evaluated is $4^{26}\epsilon$, repre-

senting a loss of about 16 decimal places of accuracy, while the correlations in Holzinger's 78-rowed matrix would need to be carried to no less than 46 places to insure even an approximate accuracy in the first decimal place of one of the regression coefficients in a formula derived by least squares for predicting one of his variates in terms of all the others.

These high limits of error may possibly be reduced in the following ways: (a) a more exact study of the error might be made by means of terms of the Taylor series of orders higher than the first; (b) the positive definite character of a correlation matrix (or other matrix of normal equations) might be utilized in an attempt to arrive at lower limits of error; (c) instead of considering the maximum possible error we might depend on some mutual cancellation of different errors and content ourselves with statements in terms of probability. The compounding of different errors of rounding, which may individually be regarded as having a probability distribution of uniform density over a fixed range, quickly gives rise to an almost exactly normal distribution of known mean and variance, so that the probability approach is attractive. However the limits of error obtained in this way with, for example, a five per cent level of probability of a greater error, though somewhat smaller than the limits associated with certainty, are disappointingly large. Investigations of the types (a) and (b) have not been made, they would apparently be very cumbersome, and (a) might have the effect of increasing the error limits considered above instead of cutting them down. Use of the check column does not provide any safeguard against the errors of rounding appearing in the original correlations, though from the probability standpoint, a carefully devised use of the check column may mitigate the accumulation of errors in successive stages.

To control such errors reliance is often placed in a substitution of the solution obtained in the given equations This is not completely satisfactory, since under some circumstances large errors in the solution may yield only slight deviations of the left from the right members of the equations, and since some deviations must be expected in any case in which only a limited number of decimals is carried along Moreover this substitution, even if it reveals the existence of errors, does not usually make clear at once what should be done about them. A recalculation to a larger number of decimal places is horribly laborious There is here a distinct need of using an iterative process for improving on the solution obtained, and setting definite limits for the errors.

4. The classical iterative method. The iterative method which seems to be the oldest and the most used for solving systems of linear equations, and which may like all other methods of doing this be applied to find the inverse of a matrix, is that of Gauss and Seidel. It seems also to be used in the "method of relaxations" [26], which has been recommended to engineers but lacks limits of error and measures of rate of convergence.

This classical method, starting with any assumed values for the unknowns, begins by changing the value for the first unknown so as to satisfy the first equation; this is possible if the coefficient is different from zero. The revised

set of trial values is then further altered by changing the second unknown so as to satisfy the second equation. Then the third unknown is altered so that the third equation will be satisfied, and so forth. When all the unknowns have been thus altered the cycle may be begun again, and repeated until the differences between consecutive values of each unknown become small enough to indicate a satisfactory convergence. The method converges if the matrix A of the coefficients a_{ij} is positive definite, as it is for the normal equations of least squares, and also in certain engineering applications [7, 8, 9]. Moreover the character of being positive definite insures that each a_{ii} differs from zero, so that the successive adjustments indicated are all actually possible. In the published discussions, proofs of convergence have sometimes been omitted, and in some cases (e.g. [30], Sec. 130) the proofs are incomplete. Even the fuller proofs [13] and [16] fail to give explicit limits for the errors in stopping at any particular stage. But from the discussion [16, pp. 502, 504] it is easy to see that positive numbers d_i and k exist, with $k < 1$, such that the error in the mth estimate of x_i is less than $d_i k^m$. This limit of error diminishes in geometric progression with successive iterations; hence the number of decimal places of accuracy increases approximately in arithmetic progression. The progression is however irregular and the trial values may fluctuate considerably. Numerical determination of limits of error does not appear to be easy. Experience with the method indicates that it is satisfactory only in case a really good approximation is available to begin with, in spite of its universal convergence.

5. An acceleration and extension of the classical iteration. This classical scheme may be improved in the following way if numerous cycles of revision of the trial values are expected to be needed for the requisite accuracy. The first step, consisting of replacing the trial value x_1 by

$$x_1' = (g_1 - a_{12}x_2 - \cdots - a_{1p}x_p)/a_{11}$$

and leaving $x_2, \cdots, x_p$ unchanged, amounts to subjecting the $p + 1$ variables $x_0, x_1, \cdots, x_p$ to the homogeneous transformation

$$\begin{aligned} x_0' &= x_0 \\ x_1' &= (g_1 x_0 - a_{12}x_2 - \cdots - a_{1p}x_p)/a_{11} \\ x_2' &= x_2 \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ x_p' &= x_p , \end{aligned}$$

where the symbol x_0, introduced for convenience in order to make these equations homogeneous, is always equal to unity. The matrix of the transformation,

$$T_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ g_1/a_{11} & 0 & -a_{12}/a_{11} & -a_{13}/a_{11} & \cdots & -a_{1p}/a_{11} \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix},$$

is of course singular. If X_0 denote the one-column, $(p+1)$-rowed matrix of the initial trial values, with unity at the head of the column, the column matrix $X_1 = T_1X_0$ is the result of this first operation, again with unity at the head of the column. The trial values obtained by the second operation appear likewise in the column matrix $X_2 = T_2X_1 = T_2T_1X_0$, where

$$T_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & 0 & \cdots & 0 \\ g_2/a_{22} & -a_{21}/a_{22} & 0 & -a_{23}/a_{22} & -a_{24}/a_{22} & \cdots & -a_{2p}/a_{22} \\ 0 & 0 & 0 & 1 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix}.$$

The result of a complete cycle of substitutions may be written $X_p = T_pT_{p-1} \cdots T_2T_1X_0$, where the matrices T_i are of the same simple character illustrated by T_1 and T_2. This same result will be obtained, because of the associative law of matrix multiplication, if we first calculate numerically the matrix

$$T = T_pT_{p-1} \cdots T_2T_1$$

and then $X_p = TX_0$. (Experience shows that computers need at this point the caution that the matrices must be arranged in their proper order. A good procedure is first to form T_2T_1, then to multiply this by T_3 on the left, etc.). This requires rather more work than the original Gauss-Seidel scheme, and therefore is not worth while if only one cycle of substitutions is needed.

The advantage lies in the fact that T may readily be squared, and T^2X_0 gives a result equivalent to that of two full cycles of iteration by the Gauss-Seidel method. Furthermore, T^2 may be squared to give T^4, which may also be squared, and so on. Obviously k such squarings give a matrix which, when multiplied by X_0, yields the same result as 2^k complete cycles of the original substitutions. In terms of the number k of squarings the number of decimal places of accuracy tends to increase in geometrical instead of arithmetic progression. This modification of the classical method does not seem to have been published heretofore, though both it and the method of Section 7 have been in use by the author and his students since 1936.

R. A. Fisher [11, Sec. 29] has introduced the valuable method of finding the inverse of a matrix A by solving together p systems, each of p equations in p unknowns, with the same matrix A of coefficients, but different columns of unknowns; these several columns of unknowns are the elements of the identical matrix. The technique of carrying this out by any of the methods resembling that of Doolittle is a simple extension involving replacement of the right-hand members of the equations by 1's and 0's and carrying along p such columns instead of one while applying exactly the same linear operations to the rows as in the older problem. This, like the problem of solving linear equations, has been elegantly adapted to efficient calculation with modern machines by Dwyer

[7, 8, 9]. The foregoing iterative methods may also be applied in this case, but the matrix T will be different for the different columns. When the given matrix is symmetric (as is implied by the positive definite character assumed in the proofs of convergence) the number of iterations required is generally cut down because the determination of each column determines also the elements of the corresponding row which lie in other columns. Iteration by groups [14] may well have a place here.

An observation of A. C. Aitken's [1] is noteworthy in connection with the solution of equations with a non-symmetric matrix, and with the finding of the inverse of such a matrix. Writing the equations in the matrix form $AX = G$, we see that the solution $X = A^{-1}G$ is also the solution of the system $(A'A)X = A'G$, where A' is the transverse (also called the transpose or conjugate) of A. Evidently $A'A$ and $A'G$ can be formed by direct multiplications and additions, without divisions. Since $A'A$ is symmetric, any of the methods for solving symmetric equations are applicable to the new system. To find the inverse of A we may first find the inverse of the symmetric matrix $A'A$ and then postmultiply it by A'; for $(A'A)^{-1}A' = A^{-1}$.

6. Roots, norms and convergence of matrices. The *norm* of a matrix A may be defined as the square root of the sum of the products of its elements by their complex conjugates, and denoted by $N(A)$. If A is real and a_{ij} is the element in the ith row and jth column,

$$N(A) = \sqrt{\Sigma\Sigma a_{ij}^2}. \tag{6.1}$$

This is the same function which Wedderburn [29, p. 125] defines as the absolute value of A and denotes by A with a heavy vertical bar on each side. Since it is rather troublesome to avoid confusing this with the determinant of A, we use the notation $N(A)$, though the analogy with the ordinary absolute value of a quantity is very suggestive in connection with proofs of convergence and limits of error obtained by means of the "triangular inequalities" below. Rella [25] gives a different definition of the absolute value of the matrix as the maximum of the absolute values of its roots.

The triangular inequalities, whose proof is easy with the help of the Cauchy inequality, are:

$$N(A + B) \leq N(A) + N(B), \tag{6.2}$$

$$N(AB) \leq N(A)N(B). \tag{6.3}$$

From the last it follows that for any positive integer m,

$$N(A^m) \leq [N(A)]^m. \tag{6.4}$$

Hence if $N(A) < 1$, the limit of $N(A^m)$ as m increases is zero. It then follows that the limit of A^m itself is zero, i.e. that each of its elements approaches zero, because of the definition of the norm.

The identical matrix of p rows, which we shall denote simply by 1, has the

norm $\sqrt{p}$, while a scalar matrix k (i e one with the quantity k in each element of the principal diagonal and zeros elsewhere) has the norm $k\sqrt{p}$. The norm of a p-rowed orthogonal matrix is $\sqrt{p}$.

The roots of a square matrix, also known as the latent roots or characteristic roots, are the values $\lambda_1, \cdots, \lambda_p$ of λ for which the determinant obtained by subtracting λ from each element of the principal diagonal vanishes. By expanding this determinant in powers of λ and using a relation between roots and coefficients of an equation, it is evident that the sum of the roots equals the sum of the elements in the principal diagonal. This sum is known as the *trace* of the matrix and denoted by $\mathrm{tr}(A)$. Thus

$$\lambda_1 + \lambda_2 + \cdots + \lambda_p = \mathrm{tr}(A). \tag{6.5}$$

From the definitions of the transverse and norm of A it is plain that

$$[\mathrm{N}(A)]^2 = \mathrm{tr}(AA') \tag{6.6}$$

if A is real.

If $f(x)$ is any polynomial in x, $f(A)$ is a matrix whose roots are known [29, p 30] to be $f(\lambda_i)$, $(i = 1, 2, \cdots, p)$. In particular, the roots of A^m are λ_i^m. Consequently

$$\lambda_1^m + \lambda_2^m + \cdots + \lambda_p^m = \mathrm{tr}(A^m). \tag{6.7}$$

All the roots of a zero matrix are zero. But the fact that all the roots of a matrix are zero does not necessarily imply that the matrix is zero; for example the roots of

$$\begin{bmatrix} 2 & -1 \\ 4 & -2 \end{bmatrix} \tag{6.8}$$

are both zero But for real symmetric matrices the vanishing of all the roots does imply the vanishing of the matrix; for the sum of the squares of the elements of a symmetric matrix equals the sum of squares of the roots, since $A = A'$, and by (6.7), (6.6) and (6 1),

$$\Sigma\lambda_i^2 = \mathrm{tr}(A^2) = \mathrm{tr}(AA') = [N(A)]^2 = \Sigma\Sigma a_{ij}^2.$$

Moreover, by continuity considerations, a sequence of p-rowed symmetric matrices must approach zero if all the roots approach zero, and conversely.

From this it is evident that a necessary and sufficient condition that A^m approach zero as m increases, when A is symmetric, is that all the roots of A be less than unity. This provides a sharper criterion of convergence than the requirement that $N(A) < 1$, which is sufficient but not necessary for convergence The latter is however far easier to apply in most numerical work, since it is far easier to compute $N(A)$ than the greatest root. Moreover it is easy to set an upper bound for $N(A)$ in various ways, of which the crudest is to notice that, by (6 1), $N(A)$ cannot exceed p times the greatest absolute value of any

element of A. Also, the test in terms of the norm is applicable to asymmetric as well as symmetric matrices.

From these considerations regarding the convergence of A^m we deduce at once the following result. If the norm of a square matrix is less than unity, then all the roots are less than unity in absolute value. The converse is not true, as the example (6.8) shows.

For any real square matrix A, symmetric or not,

$$\lambda_1^2 + \lambda_2^2 + \cdots + \lambda_p^2 \leq [N(A)]^2. \tag{6.9}$$

To prove this, we observe first that $2a_{ij}a_{ji} \leq a_{ij}^2 + a_{ji}^2$, and consequently $\mathrm{tr}(A^2) \leq \mathrm{tr}(AA')$. From (6.7) and (6.6) we then have $\Sigma\lambda_i^2 \leq \mathrm{tr}(AA') = [N(A)]^2$. This reasoning shows incidentally that $\Sigma\lambda_i^2$ is real, though the individual roots may be complex.

Not only for investigating convergence, but also in the important but neglected problems of setting definite limits of error after a finite number of steps, the norm is an extremely useful function. If a matrix is to be computed with such accuracy that the error in each element is less than δ, and A is the matrix of errors, the requisite accuracy will according to (6.1) be attained when $N(A) < \delta$. The definition and theorems regarding the norm are valid without any restriction to square matrices, for which alone the roots are defined. For example, we may use the norm to derive an inequality concerning the solution of the system of p linear equations

$$\Sigma a_{ij}x_j = g_i,$$

which may be written in matrix form $AX = G$, where A is a square matrix and X and G are matrices each of one column and p rows. From (6.3) we find $N(G) \leq N(A)N(X)$, whence

$$N(X) \geq N(G)/N(A).$$

We shall now deduce a result which seems to be new to matrix theory and which we shall later apply to find limits of error. If A is any matrix such that $1 - A$ is non-singular the identity

$$(1 - A)^{-1} = 1 + A + A^2 + \cdots + A^{m-1} + A^m(1 - A)^{-1}$$

holds, and may be demonstrated exactly as if A were an ordinary scalar quantity. Suppose that $N(A) \leq k < 1$. Taking the norm and using (6.2), (6.3) and (6.4), we have

$$N[(1 - A)^{-1}] \leq p^{1/2} + k + k^2 \cdots + k^{m-1} + k^m N[(1 - A)^{-1}].$$

Since $k < 1$ we may solve for $N[(1 - A)^{-1}]$. Summing the geometric progression, we obtain:

$$N[(1 - A)^{-1}] \leq \frac{p^{1/2} - 1}{1 - k^m} + \frac{1}{1 - k}.$$

This holds for every positive integral value of m, and therefore in the limit when m becomes infinite. Thus we find that

$$N[(1 - A)^{-1}] \leq p^{1/2} - 1 + \frac{1}{1 - k} \tag{6.10}$$

whenever $N(A) \leq k < 1$.

7. An efficient inversion procedure. Let C_0 be an approximation to the inverse of a matrix A, and consider the following sequence of operations. Calculate

$$C_1 = C_0(2 - AC_0), \tag{7.1}$$

and then in turn $C_2, C_3, \cdots$ where

$$C_{m+1} = C_m(2 - AC_m). \tag{7.2}$$

Let us inquire as to the conditions under which the sequence of matrices C_m converges to A^{-1}, the maximum error that may be committed in stopping at any stage, and the rate of convergence. Suppose that C_0 is a good enough approximation to A^{-1} to make the roots of the matrix

$$D = 1 - AC_0 \tag{7.3}$$

all less than unity in absolute value. Then increasing powers of D approach zero, and the convergence of C_m to A^{-1} will follow from the relation

$$C_m = A^{-1}(1 - D^{2^m}), \tag{7.4}$$

which will now be proved by mathematical induction. From (7.1) and (7.3),

$$C_1 = A^{-1}(AC_0)(1 + D) = A^{-1}(1 - D)(1 + D) = A^{-1}(1 - D^2),$$

so that (7 4) is verified for $m = 1$. Now assume (7.4) for a particular value of m, and substitute it in (7.2). This gives

$$C_{m+1} = A^{-1}(1 - D^{2^m})(1 + D^{2^m}) = A^{-1}(1 - D^{2^{m+1}}),$$

which being of the same form as (7.4) completes the induction.

If $N(D) \leq k < 1$ the roots of D are all less than unity in absolute value, as shown in Sec. 6, and the foregoing result holds. Assuming this to be true we now derive an upper bound for the error in C_m in terms of k and $N(C_0)$. According to (7.3),

$$A^{-1} = C_0(1 - D)^{-1}.$$

Hence, by (7.4),

$$C_m - A^{-1} = -A^{-1}D^{2^m} = -C_0(1 - D)^{-1}D^{2^m}.$$

Therefore, by (6.3), (6.4) and (6.10),

$$N(C_m - A^{-1}) \leq N(C_0)k^{2^m}\left(p^{1/2} - 1 + \frac{1}{1 - k}\right). \tag{7.5}$$

This sets an upper bound for the difference between each element of C_m and the corresponding element of A^{-1}. A slightly looser but simpler limit may be obtained from this in terms of the greatest absolute value c of any element of C_0. Since $N(C_0) \leq cp$,

$$N(C_m - A^{-1}) \leq k^{2^m} cp \left(p^{1/2} - 1 + \frac{1}{1-k}\right). \tag{7.6}$$

The great value of this method, whenever a good enough initial approximation is available to make $N(D)$ less than unity, is that the number of decimal places of sure accuracy increases in *geometric* progression, rather than in arithmetic progression as with the usual methods. Consequently this method will always be the most efficient if a sufficiently large number of decimal places is required. Moreover, a limit can be set in advance for the number of iterations that will be required in order to insure any required degree of accuracy. If certainty of correctness in the sth decimal place is required we may choose m so that the right-hand member of (7.6) is less than $10^{-s}/2$. In terms of logarithms to the base 10 the number of decimal places whose accuracy is assured by m iterations is thus at least

$$2^m \mid \log k \mid - \log 2 - \log cp[p^{1/2} - 1 + (1-k)^{-1}]. \tag{7.7}$$

These limits of error can be bettered after some iterations have actually been made. When C_r becomes available we may calculate $k_r = N(1 - AC_r)$, which may be used in place of k in the formulae just derived if m is replaced by $m - r$, and is generally enough smaller than k to make a marked improvement.

The elements of the matrix of errors will actually, of course, be smaller than the norm of this matrix in every practical case, in a ratio fluctuating about p^{-1}. The limits obtained by our formulae can be reached only in case the entire error of the matrix C_m is concentrated in one element, a very unlikely event. Thus the limits given above will usually be quite conservative.

As the iteration proceeds the elements of the matrix $D_m = 1 - AC_m = D^{2^m}$ will diminish rapidly in case of convergence. For this reason it may sometimes be better to calculate C_{m+1} not directly from (7.2), but from the formula

$$C_{m+1} = C_m + C_m D_m \tag{7.8}$$

in which the last term can be regarded as a correction of C_m which will often be very small. This method, however, lacks the self-checking feature, so that its use at the final stage is dubious.

This iterative process has been noticed previously [12, p. 120], but without a limit of error or observation of the geometric progression in the number of accurate digits.

If the initial approximation is not good enough to make $N(D) < 1$, it may be improved by other methods, such as those of Sections 4 and 5, to the point at which this more rapid method becomes applicable. But in some cases (e.g. the second example of §8) the method converges even though $N(D) > 1$, as

may be demonstrated at a later stage at which the norm of the matrix corresponding to D becomes numerically less than unity.

For the mass of least-square and other problems in which the inverse of a matrix is needed, the best procedure appears to begin with one of the methods described by Dwyer [7, 8, 9], carried to a small number of decimal places, and then to calculate D from (7.3), a step equivalent to substituting the approximate solution obtained into the equations. It may then be evident at a glance that the norm of D is so small that the method of the present section will converge rapidly to give as many more places as desired. If $N(D)$ is too large for this, and if gross errors have been eliminated, there is a choice between recalculation from the beginning, the classical iterative process, and the acceleration of this process by matrix-squaring, with perhaps some iteration by small groups. The choice will depend partly on how much the elements of D need to be reduced. The classical iteration (or sometimes the process of this section) is appropriate for correcting a slight excess of $N(D)$ over unity, its matrix-squaring extension for larger alterations.

Let E_0 be the error in C_0, so that $C_0 = A^{-1} + E_0$. Then by (7.1),

$$C_1 = (A^{-1} + E_0)(1 - AE_0) = A^{-1} - E_0 A E_0 .$$

If E_1 is the error in C_1, so that $C_1 = A^{-1} + E_1$, we thus have

$$E_1 = -E_0 A E_0 .$$

If A is symmetric, we naturally take C_0 as a symmetric matrix, and this will cause E_0, C_1, and E_1 also to be symmetric. If also A is positive definite, it will follow from the last equation that E_1 is *negative* definite, or negative semi-definite. Consequently the diagonal elements of C_1 tend to underestimate the corresponding elements of A^{-1}, and never exceed them. Furthermore, the value of a quadratic form whose matrix is A^{-1} will be at least as great as the estimate of it based on C_1. The squares, both of the multiple correlation coefficient and the generalized Student ratio [15], can be expressed as such quadratic forms. Hence both these statistics are slightly underestimated when C_1 is used in place of the true matrix of coefficients. Later approximations C_m do not change the signs of these biases, though they make their magnitudes approach zero in case the conditions for convergence are satisfied, and definite limits converging to zero are easily found for them in such cases from the results above.

8. Illustrations and further comments. We shall indicate symmetric matrices by writing only the elements on and above the principal diagonals.

To illustrate various methods Dwyer [7] has evaluated the inverse of

$$A = \begin{bmatrix} 1.0 & .4 & .5 & .6 \\ & 1.0 & .3 & .4 \\ & & 1.0 & .2 \\ & & & 1.0 \end{bmatrix}$$

as

$$\begin{bmatrix} 2.0710 & -.1913 & -.7759 & -1.0109 \\ & 1.2842 & -.2186 & -.3552 \\ & & 1.3989 & .2732 \\ & & & 1.6940 \end{bmatrix}.$$

If the accuracy of the calculation had been only such as to insure correctness in the first decimal place the approximation to A^{-1} would have been

$$C_0 = \begin{bmatrix} 2.1 & -.2 & -.8 & -1.0 \\ & 1.3 & -.2 & -.4 \\ & & 1.4 & .3 \\ & & & 1.7 \end{bmatrix}.$$

It is easy by mental arithmetic alone, without the use of a machine or side calculations, to see that

$$D = 1 - AC_0 = \begin{bmatrix} -.02 & .02 & 0 & -.01 \\ 0 & 0 & -.02 & .03 \\ .01 & -.01 & 0 & -.02 \\ -.02 & .04 & -.02 & 0 \end{bmatrix}$$

and further that $N(D) = \sqrt{.0052} = .072$. This is so much less than unity that the iteration process of §7 will converge rapidly. As a matter of fact, without determining the sum of the squares of the elements of D we could have observed at a glance that $N(D)$ must be less than four times the greatest absolute value of an element, and thus have a value less than .16. In the same way $N(C_0)$ is seen to be less than 8.4; actually it equals 3.8588. The latter value, with $k = .072$, $p = 4$, substituted in (7.5) gives for the norms of the successive error matrices $E_m = C_m - A^{-1}$,

$$\begin{aligned} N(E_0) &\leq 8.03k &&= .578, \\ N(E_1) &\leq 8.03k^2 &&= .0414, \\ N(E_2) &\leq 8.03k^4 &&= .000216, \\ N(E_3) &\leq 8.03k^8 &&= .000\,000\,0058. \end{aligned}$$

This promises merely that after one application of the iterative process the results will be accurate to one decimal place, which we know already but might not have known for sure in such a case; that a second iteration will give results accurate to three places, and that a third will give results accurate to about eight places. These estimates will however be improved after actually computing C_1. This may well be done by (7.1) if a machine is available; otherwise, and almost as easily, by (7.8) we obtain

$$C_1 = \begin{bmatrix} 2.070 & -.190 & -.776 & -1.011 \\ & 1.282 & -.218 & -.355 \\ & & 1.398 & .274 \\ & & & 1.692 \end{bmatrix}$$

and $N(C_1) = 3\,8163$. (We have now passed beyond the stage of easy mental calculation, but might alternatively use the easy upper bound 8.28 for $N(C_1)$, obtained as before.) We shall use this value instead of $N(C_0)$ in (7.5) and at the same time use for k the value of $N(D_1)$, where

$$D_1 = 1 - AC_1 = 1 - AC_0(1 + D) = 1 - (1 - D)(1 + D) = D^2.$$

This is most easily found from D, from which it may be written down directly by mental calculation:

$$D_1 = 10^{-4} \times \begin{bmatrix} 6 & -8 & -2 & 8 \\ -8 & 14 & -6 & 4 \\ 2 & -6 & 6 & -4 \\ 2 & -2 & -8 & 18 \end{bmatrix}$$

The norm of D_1 is seen by the crude method to be less than .0072, and is actually .003212. Taking the latter value for k we have, similarly to (7.5),

$$N(E_m) \leq N(C_1)k^{2^{m-1}} \times 2\,00323 = 7.645k^{2^{m-1}}$$

Thus,

$$\begin{aligned} N(E_1) &\leq .0246, \\ N(E_2) &\leq .000\,0789, \\ N(E_3) &\leq .000\,000\,000\,8. \end{aligned}$$

The reduction in these limits of error is due to the difference between $[N(D)]^2 = .0052$ and $N(D^2) = .003212$.

Using $C_2 = C_1 + C_1D_1$ we obtain:

$$C_2 = \begin{bmatrix} 2.0710366 & -.1912542 & -.7759568 & -1.0109294 \\ & 1.2841486 & -.2185780 & -.3551910 \\ & & 1.3989056 & .2732260 \\ & & & 1.6939852 \end{bmatrix}.$$

From this we calculate

$$D_2 = 1 - AC_2 = 10^{-8} \times \begin{bmatrix} 112 & -164 & -40 & 168 \\ -164 & 288 & -136 & 88 \\ 64 & -128 & 100 & -104 \\ 48 & -32 & -184 & 364 \end{bmatrix},$$

agreeing with the value obtained from the formula $D_2 = D_1^2$, and finally $C_3 = C_2(1 + D_2) =$

$$\begin{bmatrix} 2.071\,038\,458 & -.191\,256\,831 & -.775\,956\,284 & -1.010\,928\,962 \\ & 1.284\,153\,005 & -.218\,579\,235 & -.355\,191\,257 \\ & & 1.398\,907\,104 & .273\,224\,045 \\ & & & 1.693\,989\,071 \end{bmatrix}$$

which as shown above is correct to at least eight decimal places, and doubtless more, in each element. The estimate of A^{-1} obtained by Dwyer by several

direct methods to four places is corroborated by this result excepting for a slight error in the element in his first row and third column.

(ii) Suppose that the approximation in the foregoing example had been even cruder, with determination of the elements of A^{-1} only to the nearest integer. This would give

$$C_0 = \begin{bmatrix} 2 & 0 & -1 & -1 \\ & 1 & 0 & 0 \\ & & 1 & 0 \\ & & & 2 \end{bmatrix},$$

$$D = \begin{bmatrix} .1 & -.4 & .5 & -.2 \\ -.1 & 0 & .1 & -.4 \\ .2 & -.3 & .5 & .1 \\ 0 & -.4 & .4 & -.4 \end{bmatrix}.$$

The sum of the squares of the elements of D is 1.51, so that the norm is greater than unity, and it is not clear at this stage whether the iterative process we have been using will converge or not. But upon computing

$$D^2 = \begin{bmatrix} .15 & -.11 & .18 & .27 \\ .01 & .17 & -.16 & .19 \\ .15 & -.27 & .36 & .09 \\ .12 & .04 & 0 & .36 \end{bmatrix}$$

we find that $N(D^2) = \sqrt{.6093} = .7806$, and since this is less than unity we are assured that the process will converge. We may write immediately, without use of a machine or written side calculation:

$$C_1 = C_0 + C_0D = \begin{bmatrix} 2.0 & -.1 & -.9 & -1.1 \\ & 1.0 & .1 & -.4 \\ & & 1.0 & .3 \\ & & & 1.4 \end{bmatrix}.$$

Utilizing the value of D^2 already determined, we readily find

$$C_2 = C_1 + C_1D^2 = \begin{bmatrix} 2.032 & -.138 & -.848 & -1.056 \\ & 1.138 & -0.42 & -.372 \\ & & 1.182 & .274 \\ & & & 1.558 \end{bmatrix}.$$

From this point on a machine is needed for efficiency. The next step is to calculate D^4, either by squaring D^2 or by the formula $D^4 = I - AC_2$; both methods may be used as a check. The result is:

$$D^4 = 10^{-4} \times \begin{bmatrix} 808 & -730 & 1094 & 1330 \\ 20 & 786 & -830 & 890 \\ 846 & -1560 & 1998 & 540 \\ 616 & 80 & 152 & 1696 \end{bmatrix}.$$

We may now consider the accuracy of further approximations, inserting in (7.5) $N(C_2) = \sqrt{13.385572} = 3.659$ in place of $N(C_0)$, $m - 2$ for m, and $k = N(D^4) = .4119$. Thus

$$\begin{aligned} N(E_2) &\leq (9.8807)(.4119) &&= 4.0699 \\ N(E_3) &\leq (9.8807)(.4119)^2 &&= 1.6764 \\ N(E_4) &\leq (9.8807)(.4119)^4 &&= .2844 \\ N(E_5) &\leq (9.8807)(.4119)^8 &&= .00819 \\ N(E_6) &\leq (9.8807)(.4119)^{16} &&= .000\ 006\ 79. \end{aligned}$$

Because of the roughness of the initial approximation in this case the convergence is rather slow at first, but later it is much accelerated. So far as the limits found above show, five iterations are necessary to be sure of even approximate two-place accuracy in the results (somewhat better limits could be obtained after actually calculating C_2, still better ones from C_3, etc.), but the sixth iteration gives results sure to be accurate nearly to five places. Perhaps the best treatment of a numerical case of this kind is to work out the solution by Dwyer's method to two, three or four places, and then to apply the iterative process once, and as many more times as necessary to obtain the required accuracy.

The final step should, for the sake of checking, be a calculation of C_{m+1} from $C_m(2 - AC_m)$, rather than from $C_m + C_mD^{2^m}$.

Upon observing that $N(D) > 1$ we might have used the Seidel process to improve each row of C_0. This process is however extremely slow, and in the present example is markedly inferior to that used above.

(iii) If we start from the result which Dwyer gives to four decimal places as C_0, we obtain

$$D = 1 - AC_0 = 10^{-5} \times \begin{bmatrix} 1 & 4 & -3 & -2 \\ 3 & -2 & 1 & 0 \\ -3 & 3 & -1 & 1 \\ 0 & 2 & 0 & -2 \end{bmatrix}.$$

We find $N(C_0) = 3.8188$, and putting $k = N(D) = .00085$ we have from (7.5),

$$N(E_m) \leq 3.8188\ (.00085)^{2^m}(2.00085) \leq (7.6408)(.00085)^{2^m}.$$

Thus $N(E_1) \leq .000\ 0055$,

$N(E_2) \leq .000\ 000\ 000\ 0004$.

9. Certain other methods of successive approximation. A class of methods for solving linear equations, which may be extended to find the inverse of a matrix, is given by Frazer, Duncan and Collar [12, pp. 132–133], generalizing a method of J. Morris. In this method the matrix A of the coefficients in the linear equations, or the matrix to be inverted, is written as the sum of an easily inverted matrix V, for example a diagonal or triangular matrix, and another

matrix W. Then

$$A^{-1} = (1 + V^{-1}W)^{-1}V^{-1} = (1 - f)^{-1}V^{-1},$$

where $f = -V^{-1}W$. If the latent roots of f are all less than unity in absolute value, and *a fortiori* if $N(f) < 1$, the series

$$1 + f + f^2 + f^3 + \cdots = (1 - f)^{-1}$$

converges. To solve the equations $AX = G$, where X and G are column vectors (i.e. matrices of one column) is to determine

$$X = A^{-1}G = (1 - f)^{-1}H,$$

where $H = V^{-1}G$. The method of Frazer, Duncan and Collar is to calculate the successive vectors

$$X_0 = H, \qquad X_1 = H + fX_0, \qquad X_2 = H + fX_1, \cdots, X_r = H + fX_{r-1}, \cdots.$$

It is clear that

$$X_r = (1 + f + f^2 + \cdots + f^r)H.$$

The error in X_r is therefore the vector

$$E_r = f^{r+1}(1 - f)^{-1}H.$$

We may ascertain a limit for the errors if $N(f) \leq k < 1$. Indeed, by (6.3) and (6.10),

$$N(E_r) \leq k^{r+1}\left(p^{1/2} - 1 + \frac{1}{1-k}\right)N(H),$$

where p is the number of unknowns; and no individual unknown will have an error greater than $N(E_r)$.

Convergence of this method, if existent, may be accelerated by matrix-squaring. Indeed, upon calculating in turn $f^2, f^4, f^8, f^{16}, \cdots$ by repeated squarings, we need only to work with the sequence

$$X_0 = H, \qquad X_1 = (1 + f)X_0, \qquad X_3 = (1 + f^2)X_1,$$

$$X_7 = (1 + f^4)X_3, \qquad X_{15} = (1 + f^8)X_7, \cdots,$$

omitting the intermediate approximations. This will be worth while for solving a single set of equations only in case such great accuracy is required as to demand the use of rather high powers of f. Each squaring of f consists of the formation of p^2 sums of products, so that determination of, say, X_{31} by this method requires $4p^2$ such sums after f has been determined, in addition to the $5p$ involved in finding $X_1, X_3, X_7, X_{15}, X_{31}$ after the squarings. By the method of Frazer, Duncan and Collar the corresponding number of sums of products would be $31p$. Since $4p^2 + 5p < 31p$ only in case $p \leq 6$, it appears that the matrix-squaring is justified only for six or fewer unknowns unless a larger number of terms is required. Furthermore, increasingly high powers of a matrix, to be useful, need usually to be expressed with more and more significant digits.

If more than one system of equations with the same matrix A is to be solved, these methods have the advantage that the same matrix f can be used for all the vectors G of right-hand members. In such cases the value of matrix-squaring is enhanced in comparison with that in which only a single system of equations is to be solved. Determination of A^{-1} is equivalent to solving p such systems in which the several column vectors G together constitute the identical matrix. If more than p of these systems of equations are to be solved it is best to find A^{-1} and then form the various solutions $A^{-1}G$ from the columns G of right-hand members.

It is worth noticing that the matrices $1 + f$, $1 + f^2$, etc., are commutative, as are all rational functions of a single matrix. In difficult cases this may occasionally provide a useful check.

This method differs from the other iterative methods with which we are concerned in that errors of calculation are not automatically corrected by it. This is a serious disadvantage, especially for the inexperienced computer, and makes desirable the careful maintenance of a check column. On the other hand, it does not require any preliminary knowledge of the solution. Indeed, it should be classified rather with the direct than with the iterative procedures on this account.

The critical element in determining the success of this method is the possibility or impossibility of finding suitable matrices V and W, such that V^{-1} can be calculated easily, and such that the elements of $f = -V^{-1}W$ are sufficiently small to make the roots all numerically less than unity. Morris uses for V the matrix derived from A by replacing all the elements above the principal diagonal by zeros. This insures that the corresponding positions in V^{-1} are also occupied by zeros. The other elements of V^{-1} are then determined fairly easily. If the non-diagonal elements of A, which appear in W, are sufficiently small, this fact will insure small enough elements in f to make convergence rapid.

A second method, given by Frazer, Duncan and Collar, chooses for V a diagonal matrix (one having only zero elements except in the principal diagonal), or simply the unit matrix. This choice reduces the labor of inversion to a minimum. Successful convergence will take place when the non-diagonal elements of A are sufficiently small in comparison with those in the diagonal, if V is taken as the diagonal matrix containing the diagonal elements of A.

A third method which may be useful in certain cases, particularly when some but not all of the unknowns are required, is the following. Let A be partitioned:

$$A = \left[\begin{array}{c|c} a & b \\ \hline c & d \end{array}\right],$$

where a and d are square submatrices which, being of lower order than A, are more easily inverted. Let V and W be the correspondingly partitioned matrices

$$V = \left[\begin{array}{c|c} a & 0 \\ \hline 0 & d \end{array}\right], \qquad W = \left[\begin{array}{c|c} 0 & b \\ \hline c & 0 \end{array}\right].$$

Putting $s = a^{-1}b$, $t = d^{-1}c$, we have:

$$f = -\left[\begin{array}{c|c} 0 & s \\ \hline t & 0 \end{array}\right], \qquad f^2 = \left[\begin{array}{c|c} st & 0 \\ \hline 0 & ts \end{array}\right], \qquad f^4 = \left[\begin{array}{c|c} (st)^2 & 0 \\ \hline 0 & (ts)^2 \end{array}\right], \cdots.$$

If only the first q of the p unknowns are required, a and b may be taken as matrices of q rows If G_1 and H_1 consist respectively of the first q rows of G and H, and if G_2 and H_2 consist of the remaining rows, then $H_1 = a^{-1}G_1$ and $H_2 = d^{-1}G_2$ Then, in case of convergence, the first q rows of the solution are given by

$$(1 + st)[1 + (st)^2][1 + (st)^4] \cdots H_1 - s(1 + ts)[1 + (ts)^2][1 + (ts)^4] \cdots H_2.$$

Convergence to the correct values is assured here if the norm of any power of st is less than unity, as is true if and only if the absolute values of all the roots of st are less than unity. This is easily seen to be true, since as m increases

$$\lim (ts)^m = t[\lim (st)^{m-1}]s.$$

10. A simple iterative method of solving equations. An entirely different method, whose convergence is independent of the initial trial values, is the following. To solve for the column vector X the equation $AX = G$, we may start with an arbitrary column of trial values X_0 and a scalar constant h, and then for $m = 1, 2, \cdots$ calculate X_m from

$$X_m = hG + (1 - hA)X_{m-1}.$$

If X_m is equal to X_{m-1} it is obviously the desired solution. Otherwise there is an error

$$X_m - X = (hG - X) + (1 - hA)X_{m-1} = (hA - 1)X + (1 - hA)X_{m-1}$$
$$= (1 - hA)(X_{m-1} - X) = \cdots = (1 - hA)^m(X_0 - X).$$

This converges to zero as m increases provided the latent roots of $1 - hA$ are all less than unity in absolute value If A has only real roots this is equivalent to requiring that they all be between 0 and $2/h$. In particular, if A is a correlation matrix, its roots are all real and positive. Since their sum $= tr(A) = p$, where p is the number of rows, all roots of A lie between 0 and p. Consequently the process will converge in this case if $0 \leq h \leq 2/p$. It is desirable, in order to make the error diminish as fast as possible, to take h as large as is consistent with convergence. In some cases a lower limit than p will be known for the greatest root of A, and then a smaller value than $2/p$ can be taken for h. A limit of error is obviously set by

$$N(X_m - X) \leq [N(1 - hA)]^m N(X_0 - X).$$

This method can of course be applied to find the inverse matrix.

It can also be accelerated by matrix-squaring. If we put $D = 1 - hA$ we have for example,

$$X_8 = (1 + D)(1 + D^2)(1 + D^4)hG + D^8X_0 .$$

The last term will approach zero in case of convergence, and may be omitted in this type of calculation.

Thus accelerated, the method gives decimal places of accuracy increasing in geometric instead of arithmetic progression, and is remarkably simple and straightforward. It is at its best when the roots of A are known to be closely clustered about unity. A criterion of this is that $\Sigma(\lambda_i - \bar{\lambda})^2$ shall have a small value, where $\bar{\lambda}$ is the mean of the p roots λ_i. This sum of squares equals $\Sigma\lambda_i^2 - p$ for a correlation matrix A, and $\Sigma\lambda_i^2 = tr(A^2) = \Sigma\Sigma a_{ij}^2 = p + 2\sum_{i<j} a_{ij}^2$, so that

$$\sum (\lambda_i - \bar{\lambda})^2 = 2\sum_{i<j} a_{ij}^2 .$$

Smallness of this quantity is favorable not only to this iterative method but also to those of §§4 and 5.

11. Use of the characteristic equation for inversion and for finding determinants. A method differing greatly from the others is based on the Cayley-Hamilton theorem that every matrix satisfies its own characteristic equation [29, p. 23; 4, p. 296]. This is the equation

$$f(\lambda) = \begin{vmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} - \lambda \end{vmatrix}$$

$$= e_p - e_{p-1}\lambda + e_{p-2}\lambda^2 - \cdots + (-)^{p-1}e_1\lambda^{p-1} + (-)^p\lambda^p = 0,$$

where e_r $(r = 1, 2, \cdots, p)$ is the sum of the products r at a time of the roots, and also equals the sum of the r-rowed principal minors of the matrix A. Substituting A for λ, which by the Cayley-Hamilton theorem is legitimate, multiplying by A^{-1}, and transposing a term, yields

$$(11.1) \quad e_pA^{-1} = e_{p-1} - e_{p-2}A + e_{p-3}A^2 - \cdots + (-)^p e_1A^{p-2} + (-)^{p+1}A^{p-1}.$$

This equation provides a direct method of calculating A^{-1} as soon as the elementary symmetric functions e_r of the roots of $f(\lambda) = 0$ have been evaluated. This evaluation may be accomplished by means of Newton's identities [4, p. 243] connecting the elementary symmetric functions with the power-sums. If s_r is the sum of the rth powers of the roots, these formulae give:

$$\begin{aligned} e_1 &= s_1 \\ e_2 &= \tfrac{1}{2}(e_1 s_1 - s_2) \\ e_3 &= \tfrac{1}{3}(e_2 s_1 - e_1 s_2 + s_3) \\ &\cdots \cdots \cdots \cdots \\ e_p &= \frac{1}{p}(e_{p-1} s_1 - e_{p-2} s_2 + \cdots \pm s_p). \end{aligned}$$

The procedure is to calculate in turn $A^2, A^3, \cdots, A^{p-1}$, then to obtain the s's from the diagonals of these matrices, since $s_r = tr(A^r)$, then to obtain the elementary symmetric functions $e_1, \cdots, e_{p-1}$ of the roots from Newton's formulae, and to substitute these in the right-hand member of (11.1). It is then only necessary to find and divide by c_p, which equals the determinant of A. For this, and for checking the calculations, there is a choice of methods. We may find the diagonal of A^p, without troubling to compute the whole of this matrix, from the product AA^{p-1} and also, to provide a comprehensive check, from $A^{p-1}A$ or possibly from the product of two powers of A of exponents approximating $p/2$. The sum of these diagonal elements of A^p is s_p, which may be substituted in the last of the Newton formulae above with the quantities previously found to give e_p. An alternative method is to multiply A by its adjoint e_pA^{-1}, which is computed by (11.1), to obtain the determinant c_p.

The total number of multiplications, divisions, and additions is distinctly greater by this method than by efficient direct methods such as that of Dwyer [7, 9]. On the other hand, this method is straightforward and easily checked; the divisions involved are of the simplest character, consisting only of the divisions by $2, 3, \cdots, p$ in Newton's formulae and of the final division of the adjoint matrix by one number; and for large matrices it is ideally adapted for matrix multiplication by means of punched cards. A further very important advantage of this characteristic function method is that it yields considerable additional information as a by-product. Not only the determinant of the matrix but the sums s_r of the principal minors of each order r are determined. Moreover the characteristic equation, whose coefficients would be exceedingly difficult to compute directly from definitions for a large matrix, is by this method made available for the study of the latent roots, which have great interest in themselves for numerous purposes.

The characteristic function method is applicable whether A is symmetric or not. If it is symmetric, the same is true of each of the other matrices appearing in the calculation, so that it is necessary to write only about half the elements.

An illustration using a symmetric matrix has been given by M. D. Bingham [3]. In the illustration below the matrix is not symmetric and has complex double roots and non-linear elementary divisors, so that evaluation of the roots by iterative methods, though possible, would be very slow and laborious, as shown by Aitken [2]. This is indeed the same example used by Aitken in this discussion. But it should be noted that the associated latent vectors, which are determined along with the roots in the iterative processes, require the solution of sets of $p - 1$ linear equations if the roots are found directly by solving the characteristic equation.

$$\text{Let } A = \begin{bmatrix} 15 & 11 & 6 & -9 & -15 \\ 1 & 3 & 9 & -3 & -8 \\ 7 & 6 & 6 & -3 & -11 \\ 7 & 7 & 5 & -3 & -11 \\ 17 & 12 & 5 & -10 & -16 \end{bmatrix}.$$

Then

$$A^2 = \begin{bmatrix} -40 & -9 & 105 & -9 & -40 \\ -76 & -43 & 32 & 44 & 23 \\ -55 & -22 & 62 & 20 & -10 \\ -61 & -25 & 65 & 20 & -7 \\ -40 & -9 & 110 & -14 & -40 \end{bmatrix}, \quad A^3 = \begin{bmatrix} -617 & -380 & 64 & 199 & 256 \\ -260 & -189 & -316 & 355 & 280 \\ -443 & -279 & -106 & 115 & 259 \\ -464 & -300 & -136 & 139 & 202 \\ -617 & -385 & 69 & 199 & 256 \end{bmatrix},$$

$$A^4 = (A^2)^2 = \begin{bmatrix} -1342 & -978 & -2963 & 2444 & 2006 \\ 944 & 522 & -1982 & -10 & 503 \\ -358 & -333 & -2435 & 1307 & 1331 \\ -175 & -243 & -2645 & 1217 & 1355 \\ -1312 & -963 & -2978 & 2444 & 1991 \end{bmatrix} = A \cdot A^3 \text{ (check).}$$

From the diagonals of these matrices,

$$s_1 = 5, \quad s_2 = -41, \quad s_3 = -217, \quad s_4 = -17.$$

Calculating the sum of the diagonal elements only of A^5 (on a machine, without listing them separately) from AA^4 and also, as a check, from A^2A^3 we find $s_5 = 3185$. Newton's formulae then give

$$e_1 = 5, \quad e_2 = 33, \quad e_3 = 51, \quad e_4 = 135, \quad e_5 = -225,$$

the last value being that of the determinant of A. We readily find from (11.1),

$$A^{-1} = -\frac{1}{225}\begin{bmatrix} -207 & 64 & -124 & 111 & 171 \\ -315 & 30 & 195 & -180 & 270 \\ -315 & 30 & -30 & 45 & 270 \\ -225 & 75 & -75 & 0 & 225 \\ -414 & 53 & 52 & -3 & 342 \end{bmatrix}.$$

So far, all results by this method are exact, but the division by 225 introduces recurring decimals and therefore a limited validity for the form

$$A^{-1} = \begin{bmatrix} .9200 & -.2844 & .5511 & -.4933 & -.7600 \\ 1.4000 & -.1333 & -.8667 & .8000 & -1.2000 \\ 1.4000 & -.1333 & .1333 & -.2000 & -1.2000 \\ 1.0000 & -.3333 & .3333 & 0 & -1.0000 \\ 1.8400 & -.2356 & -.2311 & .0133 & -1.5200 \end{bmatrix}.$$

The characteristic equation

$$f(\lambda) = \lambda^5 - 5\lambda^4 + 33\lambda^3 - 51\lambda^2 + 135\lambda + 225 = 0$$

may in this case be solved readily, since

$$f(\lambda) = (\lambda + 1)(\lambda^2 - 3\lambda + 15)^2$$

III. Latent Roots and Vectors

12. Direct and iterative methods. If the latent roots but not the latent vectors of a matrix are desired, as for example in a preliminary study of vibra-

tions in machinery being designed, where the important question is whether any root has a positive real part, it is only necessary to find the characteristic equation and to work with it by the methods of the theory of equations. The coefficients in the characteristic equation are the sums of the r-rowed principal minors ($r = 1, 2, \cdots, p$), and are expeditiously found directly from this definition for matrices of four or fewer rows. For large matrices, however, the calculation of so many large overlapping determinants is wasteful of effort, since many virtually equivalent calculations must be done repeatedly. Indeed, calculation by determinants in a great many situations, including the solution of linear equations, is open to this objection. The methods of §11 yield the characteristic function in a manner which, for large matrices, appears to be the best available, excepting perhaps the new method of Samuelson [25a].

When, as is commonly the case, the latent vectors are desired, a straightforward calculation directly from the definitions would require not only setting up and solving the characteristic equation, but also the solution, in the case of each root, of the set of linear equations in p unknowns whose matrix is obtained from the characteristic matrix by substituting the particular root for λ. It is this solution of linear equations that aggravates greatly the computational labor when direct methods are used.

An ingenious method has been used by R. A. Fisher [11, pp. 299 ff.]. Starting with a four-rowed determinant whose elements are linear functions of an unknown θ, Fisher calculates the value of the determinant for selected values of θ, and then by interpolation using divided differences finds the largest value of θ making the determinant zero. The point of the divided difference method is that it avoids the direct calculation of the determinant for more than a few values of θ, replacing it essentially by calculation of the fourth-degree polynomial in θ from its differences and using the fact that the fourth divided differences are constant. The linear equations are then solved in a direct manner. If applied to large matrices this would be very laborious, but it compares favorably with calculation directly from definitions in the manner suggested by reading books on algebra and solid analytic geometry. But even with large matrices Fisher's method may perhaps be the best in certain cases, e.g. if all that is desired is the root of median absolute value and if this root is real, or if it is desired to find a few real roots that are close together, with numerous others greater and another numerous group less than these. This is because the iterative methods give the real roots in the order of their absolute values, beginning with the greatest, but with the possibility of obtaining them in the opposite order by first inverting the matrix. The Mallock electrical device [22] may be used to calculate determinants, and thus to apply this method.

If A and B are p-rowed matrices and B is non-singular, the determinantal equation $|A - \lambda B| = 0$ is equivalent to $|AB^{-1} - \lambda| = 0$ and also to $|B^{-1}A - \lambda| = 0$ The column vectors X, satisfying $(A - \lambda B)X = 0$ also satisfy $(B^{-1}A - \lambda)X = 0$ and the row vectors V, satisfying $V_i(A - \lambda_i B) = 0$ also satisfy $V_i(AB^{-1} - \lambda_i) = 0$. If A and B are symmetric, $V_i = X_i'$. Thus

any problem of this type is reducible to that of finding latent roots and vectors, upon calculating B^{-1} by any method and multiplying in either order by A.

The fundamental iterative method for finding latent roots and vectors of A begins with an arbitrary matrix X_0 of a single column. This column vector is premultiplied by A to obtain a new column vector X_1. If, as is possible though unlikely, the elements of X_1 are proportional to those of X_0, they constitute one of the latent vectors of A, and the factor of proportionality is the corresponding root, for then X_0 and X_1 are solutions of the matrix equation $(A - \lambda)X = 0$. It should be observed that the latent vector is determined only to within an arbitrary scalar factor of proportionality, though we may sometimes find it convenient to normalize the vector by choosing the factor in such a way that the sum of the squares of the elements, which equals the square of the norm, is unity.

If X_1 is not proportional to X_0, the operation may be repeated by calculating $X_2 = AX_1$, then $X_3 = AX_2$, and so on. If these vectors are then normalized, or if they are divided by, say, their respective first elements, then the other elements will (in the cases of greatest practical importance) gradually approach stable values which will determine one of the latent vectors, while the successive factors of proportionality will approach the corresponding root. The convergence of this process is however apt to be rather slow. Fortunately there are several known ways of accelerating it.

Matrix-squaring is the first of these methods of accelerating convergence [17, 19]. It is clear that $X_t = A^t X_0$. Consequently one application of the iterative process with A^t is equivalent to t iterations with A. It is relatively easy to square A, and then by repeated squarings to form A^4, A^8, A^{16}, etc. The economical limit of this process is determined partly by the necessity of retaining more and more digits in the successively higher powers, but up to a point not yet determined exactly it presents very great advantages. For proceeding to the determination of latent roots of other than the maximum absolute value, with their associated vectors, this method lends itself to further shortcuts [17, 2], which seem to give it an advantage over an older method [13].

Another method of accelerating convergence, introduced by A. C. Aitken, and referred to by him as the δ^2-method, uses the ratio $\phi(t)$ of an element of X_{t+1} to the corresponding element of X_t in the function

$$\frac{\phi(t+1)\phi(t-1) - [\phi(t)]^2}{\phi(t+1) - 2\phi(t) + \phi(t-1)},$$

which converges rapidly toward the root λ_1 of greatest absolute value. If a constant c is subtracted from all three of the quantities $\phi(t+1)$, $\phi(t)$ and $\phi(t-1)$ before computing the foregoing function the result is unchanged. This fact reduces greatly the computational labor, since the decimal places of λ_1 already determined are common to all three.

If A is symmetric and we form the scalar products of $X_t = A^t X_0$ with itself

and with X_{t+1} we have

$$X_t'X_t = X_0'A^{2t}X_0, \qquad X_{t+1}'X_t = X_0'A^{2t+1}X_0.$$

The ratio of these two scalars gives an estimate of λ_1 which on the basis of the ratios of consecutive elements in a given place in the trial vectors would not be reached until a later stage of convergence, corresponding in fact to twice as many iterations. Aitken has pointed out the great value of this procedure for finding the root (but not the latent vector), and has extended the idea to asymmetric matrices, where there is a complication because of the existence of two latent vectors for each root, one determined by premultiplying by A, the other by A'.

The comprehensive paper [2] of Aitken gives an extremely valuable account of the whole problem and processes of finding the latent roots and vectors, including a survey of the various cases arising when there are multiple roots, complex roots, and non-linear elementary divisors. This paper should be studied carefully by anyone with any substantial numerical problem of this kind.

A method using rotations of two variables at a time has been devised by T. L. Kelley [21].

The remainder of this paper will be concerned with some results, believed to be new, by which useful upper limits can be set for the errors of the results yielded by iteration for latent roots and vectors of a symmetric matrix. To find such limits of error for asymmetric matrices appears to be a much more difficult and as yet unsolved problem.

13. Accuracy of iteration with symmetric matrices. If A is symmetric, as it is in most statistical problems (though with some exceptions, as in [18]), the roots are all real and the elementary divisors are linear. Moreover there exist an orthogonal matrix H and a diagonal matrix

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots \\ 0 & \lambda_2 & 0 & \cdots \\ 0 & 0 & \lambda_3 & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

such that

$$A = H\Lambda H'. \tag{13.1}$$

Since H is orthogonal, $HH' = 1$, and therefore

$$\Lambda = H'AH, \qquad A^t = H\Lambda^t H'. \tag{13.2}$$

We may associate with the successive trial vectors $X_t = AX_{t-1} = A^tX_0$ the vectors $Y_t = H'X_t$; then $X_t = HY_t$. From these equations and the second of (13.2) it is clear that

$$Y_t = H'X_t = H'A^tX_0 = \Lambda^tH'X_0 = \Lambda^tY_0.$$

Hence, if the elements of Y_0 are $y_1, \cdots, y_p$,

$$Y_t = \begin{bmatrix} y_1\lambda_1{}^t \\ y_2\lambda_2 \\ \cdot \\ \cdot \\ \cdot \\ y_p\lambda_p{}^t \end{bmatrix}.$$

Now let α_{kt} be the scalar product of X_t and X_{t-k}:

$$\alpha_{kt} = X_t'X_{t-k} = Y_t'Y_{t-k} = \Sigma y_i^2\lambda_i^{2t-k}; \tag{13.3}$$

and let

$$\nu_{kt} = \frac{\alpha_{kt}}{\alpha_{0t}} = \frac{\Sigma y_i^2\lambda_i^{2t-k}}{\Sigma y_i^2\lambda_i^{2t}}. \tag{13.4}$$

If A has a negative root this fact will become evident after a certain stage in the iteration used to obtain this root by an alternation of sign of the numbers in any one position in consecutive trial vectors. However A^2, which as pointed out in §12 may well be calculated anyhow, has only positive roots, which are the squares of the roots of A, and has the same latent vectors as A. Hence we shall have results of sufficient generality for real symmetric matrices if we assume that all roots of the matrix with which we work are positive or zero, i.e. that it is positive definite or semi-definite. Let us choose the notation so that

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0.$$

Then if $k \geq 0$,

$$\alpha_{0t} = \Sigma y_i^2\lambda_i^{2t} \leq \lambda_1^k\Sigma y_i^2\lambda_i^{2t-k} = \lambda_1^k\alpha_{kt}.$$

Hence, by (13.4),

$$\lambda_1 \geq [\nu_{kt}]^{-1/k}. \tag{13.5}$$

It is known [23] that if $a_1, \cdots, a_p, c_1, \cdots, c_p$ are any positive numbers, the function

$$\left(\frac{c_1a_1^k + \cdots + c_pa_p^k}{c_1 + \cdots + c_p}\right)^{1/k}$$

increases monotonically with k. Putting $c_i = y_i^2\lambda_i^{2t}$, $a_i = \lambda^{-k}$ if $\lambda_i \neq 0$, and $c_i = a_i = 0$ if $\lambda_i = 0$, we find that the right-hand member of (13.5) decreases monotonically as k increases. Hence the best of these lower bounds for λ_1 is that corresponding to the least value of k that can be used, namely $k = 1$. Consequently the lower bound to be recommended for λ_1 is given by

$$\lambda_1 \geq \frac{1}{\nu_{1t}}. \tag{13.6}$$

From (13.4) it is easily seen that this lower bound approaches λ_1 when t increases, provided $y_1 \neq 0$.

An upper bound for λ_1 is available from the fact that the sum of the tth powers of the roots is the trace of A^t. Since we assume all $\lambda_i \geq 0$ this gives

$$\lambda_1 \leq (\operatorname{tr} A^t)^{1/t}.$$

That this upper limit converges to λ_1 when t increases is easily seen from (6.7) upon consideration of $\log (\Sigma\lambda_i^t)^{1/t}$.

A lower limit alternative to that of (13.6) is also available from $\operatorname{tr} (A^t)$, and likewise converges to λ_1. Indeed, since λ_1 is the greatest root, we have

$$\lambda_1 \geq (\operatorname{tr} A^t/p)^{1/t}.$$

We now seek limits of accuracy for the latent vector corresponding to λ_1 and estimated by X_t. If we call this vector X, and define $Y = H^{-1}X$, then $\lim Y_t^* = Y^*$, where Y^* is the normalized form of Y. Y^* has as its ith element

$$\lim_{t\to\infty} \frac{y_i\lambda_i^t}{\sqrt{\Sigma y_j^2\lambda_j^{2t}}}.$$

If $y_1 \neq 0$, and $\lambda_1 > \lambda_2 \geq \lambda_3$, this limit is ± 1 if $i = 1$, and is otherwise 0. We take the value of Y to be

$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ \cdot \\ \cdot \\ 0 \end{bmatrix}$$

in this case. If λ_1 is a multiple root, the limit of X_t^* will depend on the initial values y_i.

A useful measure of the closeness of approach of X_t to X is the "correlation coefficient" $r_t = X^{*\prime}X_t^* = Y^{*\prime}Y_t^* = \dfrac{y_1\lambda_1^t}{\sqrt{\Sigma y_j^2\lambda_j^{2t}}}$, which obviously approaches unity as t increases if $y_1 \neq 0$ and λ_1 is a *simple* root, or if $\lambda_1 = \lambda_2 = \cdots = \lambda_s > \lambda_{s+1}$ and we arrange our definitions so that $y_1 \neq 0$ and $y_2 = \cdots = y_s = 0$. In terms of the notation previously introduced, $r_t = \dfrac{y_1\lambda_1^t}{\sqrt{\alpha_{ot}}}$. The sum of the squares of the differences of corresponding elements of the normalized vectors X^* and X_t, i.e. $[N(X^* - X_t^*)]^2$, is $2(1 - r_t)$, and therefore approaches zero as r_t approaches unity. We shall seek for r_t a lower limit approaching unity as t increases.

Let us now put

$$w_{it} = \frac{y_i^2\lambda_i^{2t}}{\Sigma y_j^2\lambda_j^{2t}} = \frac{y_i^2\lambda_i^{2t}}{\alpha_{ot}}. \qquad \text{Then } r_t^2 = w_{1t} \text{ and } \sum_{i=1}^{p} w_{it} = 1.$$

For $k \geq 1$,

$$\alpha_{kt} = \Sigma y_i^2 \lambda_i^{2t-k} \geq y_2^2 \lambda_2^{2t-k} + \cdots + y_p^2 \lambda_p^{2t-k} \geq \lambda_1^{-k}(y_2^2 \lambda_2^{2t} + \cdots + y_p^2 \lambda_p^{2t})$$
$$= \lambda_1^{-k} \alpha_{ot}(w_{2t} + \cdots + w_{pt}) = \lambda_1^{-k} \alpha_{ot}(1 - w_{1t}) = \lambda_1^{-k} \alpha_{ot}(1 - r_t).$$

$$\therefore r_t^2 \geq 1 - \frac{\lambda_1^k \alpha_{kt}}{\alpha_{ot}} = 1 - \nu_{kt} \lambda_1^k.$$

This unfortunately is not a very useful lower bound for r_t, since it approaches *zero*, not unity, as t increases.

A more satisfactory result is obtained as follows. Let $\eta_i = \lambda_i^{-1}$. Then $\nu_{kt} = \sum_{i=1}^{p} w_{it} \eta_i^k$. For any value of t we may consider a distribution of a variate taking the positive values $\eta_1, \cdots, \eta_p$ with the positive weights, or probabilities, w_{it}. The kth moment of this distribution about 0 is ν_{kt}. In particular the first moment is ν_{1t}, and is evidently at least equal to η_1, which is the least of the η_i. The standard deviation is $\sigma = \sqrt{\nu_{2t} - \nu_{1t}^2}$. As t increases, ν_{1t} will approach η_1 and σ will approach zero. Hence, if $\lambda_1 > \lambda_2$, a stage will eventually be reached at which $\nu_{1t} < \eta_2$. Let

$$k = \frac{\eta_2 - \nu_{1t}}{\sigma}.$$

By the Tchebychef-Bienaymé inequality,

$$w_{2t} + \cdots + w_{pt} \leq \frac{1}{k^2},$$

and therefore

$$r_t^2 = w_{1t} \geq 1 - \frac{\nu_{2t} - \nu_{1t}^2}{(\eta_2 - \nu_{1t})^2},$$

provided t is large enough so that $\nu_{1t} < \eta_2$. This lower bound approaches unity, as desired, when t increases.

If $\lambda_1 = \lambda_2, \cdots = \lambda_k > \lambda_{k+1}$, the same proof shows that

$$w_1^{(+)} + \cdots + w_k^{(+)} \geq 1 - \frac{\nu_{2t} - \nu_{1t}^2}{(\eta_{r+1} - \nu_{1t})^2},$$

provided $\nu_{1t} < \eta_{r+1}$. The left member is the correlation of X_t with that one of the k-parameter family of latent vectors corresponding to the multiple root for which the correlation is a maximum.

In order to utilize these results we need a lower bound for η_2, or for η_{r+1}. In case $\lambda_1 \neq \lambda_2$ this requires an upper bound for λ_2. Such an upper bound may be found at the next stage through working with the reduced or "deflated" matrix used in [17] This is $A_1 = A - \lambda_1 XX'$, where X is the normalized latent vector corresponding to λ_1; and $\lambda_2^t \leq \text{tr}\,(A_1^t)$.

Since we have arrived at a definite lower limit for r_t which approaches unity

as the iterative process proceeds, and since we have found for λ_1 upper and lower bounds converging to it, a solution has been found for the troublesome problem of the degree of accuracy in stopping at any stage of the iteration for finding the greatest root and the associated latent vector. It would be possible to go on to find from these results appropriate inequalities for A_1, and then by repetition of the above arguments, for λ_2 and the second latent vector; and then likewise for the second reduced matrix A_2 and the further roots, vectors, and reduced matrices in this cyclic order. These steps may well be taken by the computer who has mastered the above argument in connection with a numerical example.

BIBLIOGRAPHY

[1] A. C. Aitken, "Studies in practical mathematics. I. The evaluation, with applications, of a certain triple matrix product," *Proc. Roy Soc. Edinburgh*, Vol. 57 (1937), pp. 172–181.

[2] A. C. Aitken, "Studies in practical mathematics, II. The evaluation of the latent roots and latent vectors of a matrix," *Proc. Roy. Soc. Edinburgh*, Vol. 57 (1937), pp 269–304.

[3] M D. Bingham, "A new method for obtaining the inverse matrix," *Jour. Amer. Stat. Assoc.*, Vol. 36 (1941), pp. 530–534.

[4] M. Bôcher, *Introduction to Higher Algebra*, New York, 1907, 321 pp.

[5] R. C. Bose and S. N. Roy, "The distribution of the Studentized D^2-statistic," *Sankhya*, Vol. 4 (1938), pp. 19–38.

[6] Prescott D. Crout "A short method for evaluating determinants and solving systems of linear equations with real or complex coefficients." *Trans. Am. Inst. Elec. Engineers*, Vol. 60 (1941).

[7] Paul S. Dwyer, "The solution of simultaneous equations," *Psychometrika*, Vol. 6 (1941), pp. 101–129.

[8] Paul S. Dwyer, "The Doolittle technique," *Annals of Math. Stat.*, Vol. 12 (1941), pp. 449–458.

[9] Paul S. Dwyer, "Recent developments in correlation technique," *Jour. Amer. Stat. Assoc.*, Vol. 37 (1942), pp. 441–460.

[10] R A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*," Vol. 7 (1936), pp. 179–188.

[11] R. A. Fisher, "*Statistical Methods for Research Workers*, London, 7th ed., 1938, 256 pp.

[12] R. A. Frazer, W. J Duncan and A. R. Collar, *Elementary matrices and some applications to dynamics and differential equations*, Cambridge Univ. Press, 1938. 416 pp.

[13] R von Mises and Hilda Pollaczek-Geiringer, "Zusammenfassende Berichte. Praktische Verfahren der Gleichungsauflösung." *Zeitschrift für angewandte Math und Mechanik*, Vol. 9 (1929), pp. 58–77 and 152–164.

[14] Hilda Geiringer, "On the numerical solution of linear problems by group iteration," *Bull. Amer. Math. Soc.*, Vol. 48 (1942), p. 370

[15] Harold Hotelling, "The generalization of Student's ratio," *Annals of Math. Stat.*, Vol. 2 (1931), pp. 360–378.

[16] Harold Hotelling, "Analysis of a complex of statistical variables into principal components," *Jour. Educ. Psych.*, Vol. 24 (1933), pp. 417–441 and 498–520. (Reprints available only from publishers, Warwick and York, Baltimore.)

[17] Harold Hotelling, "Simplified calculation of principal components," *Psychometrika*, Vol. 1 (1936), pp 27–35

[18] Harold Hotelling, "Relations between two sets of variates," *Biometrika*, Vol. 28 (1936), pp. 321–377.

[19] R. C. J. Howland, "Note on a type of determinantal equation," *Phil. Mag.*, Series 7, Vol. 6 (1928), pp. 839-842.

[20] Truman L. Kelley and Frank S. Salisbury, "An iteration method for determining multiple correlation constants," *Jour. Amer. Stat Assn.*, Vol. 21 (1920), pp. 282 292

[21] Truman L. Kelley, *Essential traits of mental life*, Harvard, 1935, 145 pp. Chap. 1.

[22] *The Mallock electrical calculating machine.* Reprint from *Engineering* (London), June 22, 1934. 8 pp.

[23] Nilan Norris, "Inequalities among averages," *Annals of Math. Stat.*, Vol. 6 (1935), pp. 27-29.

[24] Louis A. Pipes, "Matrix theory of oscillatory networks," *Jour. Applied Physics*, Vol 10 (1939), pp. 849-860

[25] T. Rella, "*Über den absoluten Betrag von Matrizen*," International Congress of Mathematicians at Oslo, 1936 (derives λ_1 as the absolute value from 5 postulates, which are proved independent).

[25a] P. A. Samuelson, "A method of determining explicitly the coefficients of the characteristic equation," *Annals of Math. Stat*, Vol. 13 (1942), pp. 424-429.

[26] R V. Southwell, *Relaxation methods in engineering science, a treatise on approximate computation*, Oxford Univ. Press, 1940. 252 pp.

[27] G Temple, "The general theory of relaxation methods applied to linear systems," *Proc. Roy. Soc. London*, Vol. 169 A (1939), pp. 476-500.

[28] A. Wald, "The classification of an individual into one of two groups." (Unpublished).

[29] J. H. M. Wedderburn, *Lectures on matrices*, New York, 1934. 200 pp.

[30] E. T. Whittaker and G. Robinson, *The calculus of observations*, London, 1924. 395 pp.

[31] S. S Wilks, "Certain generalizations in the analysis of variance," *Biometrika*, Vol. 24 (1932), pp 471-494.

Of interest in relation to the subject of this paper, though not mentioned in the text, are the following.

Arnold Dresden, "On the iteration of linear homogeneous transformations," *Bull. Amer. Math Soc.*, Vol. 48 (1942), pp. 577-579 and 949.

Paul Horst, "A method for determining the coefficients of a characteristic equation," *Annals of Math Stat.*, Vol 6 (1935), pp. 83-84.

A T. Lonseth, "Systems of linear equations with coefficients subject to error," *Annals of Math Stat.*, Vol. 13 (1942), pp 332-337.

Rufus Oldenburger "Infinite powers of matrices and characteristic roots," *Duke Math. Jour.*, Vol. 6 (1940), pp 357-361.

ON SOLUTIONS OF THE BEHRENS-FISHER PROBLEM, BASED ON THE t-DISTRIBUTION

By Henry Scheffé

Princeton University

1. The Problem. The problem [1, 2] is the interval estimation[1] of the difference of the means of two normal populations when the ratio of the variances of the populations is unknown. The reader who wishes to see the present solution before considering theoretical details will find it recapitulated in the Summary at the end and will want to refer to the following notation: $(x_1, x_2, \cdots, x_m)$ and $(y_1, y_2, \cdots, y_n)$ are random samples from normal populations with means α and β, and variances μ and ν, respectively. Define $\delta = \alpha - \beta$. We assume $m \leqq n$, and that the variates in each sample are in the order of observation, or else have been randomized.

Recently Neyman [3] has called attention to a solution which we shall designate as (B), and which is a special case of an unpublished solution of Bartlett[2]. It will be simpler to describe (B) later, but we mention now that it has the following advantages: (*i*) its validity does not depend on the values of unknown parameters, (*ii*) the required computations are simple, and (*iii*) only existing tables are needed,- the widely available Fisher t-tables. An unsatisfactory aspect of (B) is that when the sample sizes are unequal, $n - m$ of the variates y_i are completely discarded. The solution below shares with (B) the advantages (*i*), (*ii*), (*iii*); indeed, it is identical with (B) when $n = m$, but when $n \neq m$ it is free from the above objection.

2. Simple Solution. We begin with a simple restricted approach; later we will review the result from a somewhat broader standpoint. If random variables $d_1, d_2, \cdots, d_m$ are independently normally distributed with mean δ and variance σ^2, and if L and Q are defined from

$$L = \sum_{i=1}^{m} d_i/m, \qquad Q = \sum_{i=1}^{m} (d_i - L)^2,$$

then $m^{\frac{1}{2}}(L - \delta)/\sigma$ and Q/σ^2 are independently distributed; the former is a normal variable with zero mean and unit variance; the latter, $Q/\sigma^2 = \chi^2_{m-1}$, where χ^2_k is a generic notation for a random variable distributed according to the χ^2-law with k degrees of freedom. The quotient

[1] We treat the problem from the standpoint of confidence intervals, rather than significance tests, since when the former are available for δ so is a whole class of the latter, namely for any hypothesis $\delta = \delta_0$, for all δ_0. Furthermore, questions of the existence of "best" tests and "best" confidence intervals are closely related [5a].

[2] How far Bartlett followed the path of this paper is not clear from the brief mention of his results by Welch [4], except that he did establish the sufficiency of certain orthogonality conditions.

$$m^{\frac{1}{2}}(L - \delta)/[Q/(m-1)]^{\frac{1}{2}} = t_{m-1},$$

where t_k denotes generically a variable having the t-distribution with k degrees of freedom. Define $t_{k,\epsilon}$ from

$$\Pr(-t_{k,\epsilon} \leqq t_k \leqq t_{k,\epsilon}) = \epsilon. \tag{1}$$

Then a set of confidence intervals for δ with confidence coefficient ϵ is

$$|\delta - L| \leqq t_{m-1,\epsilon}\{Q/[m(m-1)]\}^{\frac{1}{2}}. \tag{2}$$

Denote by $E(l)$ the expected length of the confidence interval (2),

$$E(l) = 2t_{m-1,\epsilon}[m(m-1)]^{-\frac{1}{2}}\sigma E[(Q/\sigma^2)^{\frac{1}{2}}],$$

$$E(l) = t_{m-1,\epsilon}c_{m-1}\sigma/m^{\frac{1}{2}}, \tag{3}$$

where

$$c_k = 2k^{-\frac{1}{2}}E(\chi_k) = (8/k)^{\frac{1}{2}}\Gamma(\tfrac{1}{2}k + \tfrac{1}{2})/\Gamma(\tfrac{1}{2}k).$$

The symmetrical choice (1) of the limits on the t-distribution minimizes (3).

We consider using in connection with the confidence intervals (2) linear functions

$$d_i = x_i - \sum_{j=1}^{n} c_{ij}y_j, \qquad i = 1, 2, \cdots, m. \tag{4}$$

The variables d_i have a multivariate normal distribution. Necessary and sufficient conditions that the d_i all have the same mean δ, equal variances σ^2, and zero covariances, are easily found to be

$$\sum_{j=1}^{n} c_{ij} = 1, \qquad \sum_{k=1}^{n} c_{ik}c_{jk} = c^2\delta_{ij}, \tag{5}$$

where $\delta_{ii} = 1$, $\delta_{ij} = 0$ if $i \neq j$. If (4) are used in (2), $E(l)$ is given by (3) with $\sigma^2 = \mu + c^2\nu$. Hence to minimize $E(l)$ we must find an $m \times n$ matrix $C = (c_{ij})$ satisfying (5), and for which c^2 is minimum. The minimum value of c^2 is m/n: this is easily proved by the use of vector algebra

Let γ_i be the i-th row of C, and let ψ be the $1 \times n$ matrix $(1, 1, \cdots, 1)$. Denote the transpose of a matrix by a prime. Then the conditions (5) read

$$\gamma_i\psi' = 1, \qquad \gamma_i\gamma_j' = c^2\delta_{ij}. \tag{6}$$

First suppose vectors $\gamma_1, \gamma_2, \cdots, \gamma_m$ satisfy (6). Then it is possible to adjoin $n - m$ orthogonal vectors $\gamma_{m+1}, \cdots, \gamma_n$, so that the complete set satisfies the second group of conditions (6) Since this set is a basis in n-space,

$$\psi = \sum_{k=1}^{n} g_k\gamma_k,$$

where the g_k are scalars. Now

$$1 = \gamma_i\psi' = \sum_{k=1}^{n} g_k\gamma_i\gamma_k' = g_i c^2, \qquad i = 1, 2, \cdots, m,$$

and thus $g_1 = g_2 = \cdots = g_m = c^{-2}$. But

$$n = \psi\psi' = \sum_{k=1}^{n} g_k^2\gamma_k\gamma_k' = mc^{-2} + \sum_{k=m+1}^{n} g_k^2 c^2 \geqq mc^{-2},$$

and hence $c^2 \geqq m/n$. On the other hand this lower bound for c^2 may be attained by taking any set $\beta_1, \beta_2, \cdots, \beta_m$ of orthogonal vectors with norms m/n, that is, $\beta_i\beta_j' = \delta_{ij}m/n$, and rotating them so that their equal angles vector $\lambda = (n/m)(\beta_1 + \beta_2 + \cdots + \beta_m)$ coincides with ψ. Then $\lambda S = \psi$, where $SS' = I$. For $\gamma_i = \beta_i S$,

$$\gamma_i\psi' = \beta_i SS'\lambda' = \beta_i\lambda' = 1,$$
$$\gamma_i\gamma_j' = \beta_i SS'\beta_j' = \beta_i\beta_j' = \delta_{ij}m/n,$$

so that equations (6) are satisfied with $c^2 = m/n$.

An especially neat solution of this minimum problem was obtained by the above method; its validity may easily be verified directly. It is

$$c_{ij} = \begin{cases} \delta_{ij}(m/n)^{\frac{1}{2}} - (mn)^{-\frac{1}{2}} + 1/n, & j \leqq m, \\ 1/n, & j > m. \end{cases}$$

Then

$$d_i = x_i - (m/n)^{\frac{1}{2}}y_i + (mn)^{-\frac{1}{2}}\sum_{j=1}^{m} y_j - \sum_{j=1}^{n} y_j/n,$$

and L and Q become simply $L = \bar{x} - \bar{y}$ and

$$Q = \sum_{i=1}^{m} (u_i - \bar{u})^2, \tag{7}$$

where

$$\bar{x} = \sum_{i=1}^{m} x_i/m, \qquad \bar{y} = \sum_{i=1}^{n} y_i/n, \qquad u_i = x_i - (m/n)^{\frac{1}{2}}y_i, \qquad \bar{u} = \sum_{i=1}^{m} u_i/m. \tag{8}$$

We may now write (2) as[3]

$$\bar{x} - \bar{y} - t_{m-1,\epsilon}\{Q/[m(m-1)]\}^{\frac{1}{2}} \leqq \delta \leqq \bar{x} - \bar{y} + t_{m-1,\epsilon}\{Q/[m(m-1)]\}^{\frac{1}{2}}. \tag{9}$$

The solution (B) mentioned at the beginning, consists of taking $c_{ij} = \delta_{ij}$ in (4), so that the conditions (5) are satisfied with $c^2 = 1$. Hence for both (B) and (9) the expected length of the confidence interval is given by (3), but with $\sigma^2 = \mu + \nu$ for (B), while $\sigma^2 = \mu + (m/n)\nu$ for (9).

[3] Obvious modifications of (9) will make it suitable for "one-sided" estimation.

3. More General Solutions. We now generalize our approach to the following extent: Let L be a linear form and Q a quadratic form in the variates $x_1, \cdots, x_m, y_1, \cdots, y_n$, with coefficients independent of the parameters (i of p.) If for some constant h i. of p, and some function f of the parameters, $h(L - \delta)/f$ and Q/f^2 are independently distributed, the former according to the normal law with zero mean and unit variance, the latter according to the χ^2-law with $k - 1$ degrees of freedom, then the quotient

$$h(L - \delta)/[Q/(k-1)]^{\frac{1}{2}} \tag{10}$$

will have the t-distribution with $k - 1$ degrees of freedom, no matter what the values of the parameters.

We note that necessarily then

$$E(L) = \delta, \tag{11}$$

$$f^2 = h^2 E[(L - \delta)^2]. \tag{12}$$

The t-distribution of (10) leads to the confidence intervals

$$|\delta - L| \leqq t_{k-1,\epsilon}[Q/(k-1)]^{\frac{1}{2}}/h, \tag{13}$$

where $t_{k-1,\epsilon}$ is defined by (1), and the confidence coefficient is ϵ. Proceeding as toward (3), we find that the expected length of (13) is

$$E(l) = t_{k-1,\epsilon} c_{k-1} f/h. \tag{14}$$

If

$$L = \sum_{i=1}^{m} a_i x_i - \sum_{i=1}^{n} b_i y_i,$$

$$E(L) = \alpha \sum_{i=1}^{m} a_i - \beta \sum_{i=1}^{n} b_i. \tag{15}$$

Since a_i, b_i are i. of p., it follows from (11) and (15) that

$$\sum_{i=1}^{m} a_i = \sum_{i=1}^{n} b_i = 1. \tag{16}$$

Writing $\qquad \xi_i = x_i - \alpha, \qquad \eta_i = y_i - \beta,$

$$L - \delta = \sum_{i=1}^{m} a_i \xi_i - \sum_{i=1}^{n} b_i \eta_i, \tag{17}$$

$$E[(L - \delta)^2] = \mu \sum_{i=1}^{m} a_i^2 + \nu \sum_{i=1}^{n} b_i^2 = f^2/h^2 \tag{18}$$

from (12); thus (14) may be written

$$E(l) = t_{k-1,\epsilon} c_{k-1} \left[\mu \sum_{i=1}^{m} a_i^2 + \nu \sum_{i=1}^{n} b_i^2 \right]^{\frac{1}{2}}. \tag{19}$$

From (18) we also have

$$f^2 = a^2 \mu + b^2 \nu, \tag{20}$$

where

$$a^2 = h^2 \sum_{i=1}^{m} a_i^2, \qquad b^2 = h^2 \sum_{i=1}^{n} b_i^2$$

are i. of p.

4. Lemma. We propose to prove next that the maximum value of k is m, that is to say, it is impossible to obtain a t-distribution for a quotient (10) with more than $m - 1$ degrees of freedom. For this we need a lemma to the effect that certain well known sufficient conditions for a quadratic form to have a χ^2-distribution are also necessary.

Since under our assumptions $h^2(L - \delta)^2/f^2 = \chi_1^2$ and $Q/f^2 = \chi_{k-1}^2$ are independent, therefore $Q^*/f^2 = \chi_k^2$, where

$$Q^* = h^2(L - \delta)^2 + Q.$$

To shorten the notation, write

$$z_i = \begin{cases} x_i, & i = 1, 2, \cdots, m, \\ y_{i-m}, & i = m + 1, \cdots, m + n, \end{cases}$$

$$\alpha_i = E(z_i), \qquad \zeta_i = z_i - \alpha_i, \qquad \sigma_i^2 = E(\zeta_i^2).$$

Let

$$Q = \sum_{s,t} q_{st} z_s z_t,$$

where the indices s and t range from 1 to $m + n$ throughout. Then q_{st} is i. of p., and

$$Q = \sum_{s,t} q_{st} \zeta_s \zeta_t + 2 \sum_s q_s \zeta_s + q,$$

where

$$q_s = \sum_t q_{st} \alpha_t, \qquad q = \sum_s q_s \alpha_s.$$

From (17)

$$h^2(L - \delta)^2 = \sum_{s,t} p_{st} \zeta_s \zeta_t,$$

where p_{st} are i. of p. Putting $q_{st}^* = q_{st} + p_{st}$, q_{st}^* are i. of p., and

$$Q^* = \sum_{s,t} q_{st}^* \zeta_s \zeta_t + 2 \sum_s q_s \zeta_s + q. \tag{21}$$

The moment-generating function of Q^*/f^2 is

$$\phi(\theta) = E[\exp(\theta Q^*/f^2)] = C_1 \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \exp\left(\theta Q^*/f^2 - \tfrac{1}{2} \sum_s \zeta_s^2/\sigma_s^2\right) \prod_s d\zeta_s.$$

There exists a non-singular linear transformation from the ζ's to v's such that

$$\sum_s \zeta_s^2/\sigma_s^2 = \sum_s v_s^2 ,$$

$$\sum_{s,t} q_{st}^* \zeta_s \zeta_t = \sum_s \lambda_s v_s^2 . \tag{22}$$

Then

$$\sum_s q_s \zeta_s = \sum_s p_s v_s , \tag{23}$$

$$\begin{aligned}\phi(\theta) &= C_2 e^{\theta q/f^2} \prod_s \int_{-\infty}^{+\infty} \exp\{-\tfrac{1}{2}[v^2 - 2\theta(\lambda_s v^2 + 2p_s v)/f^2]\}\, dv \\ &= e^{\theta q/f^2} \prod_s (1 - 2\theta\lambda_s/f^2)^{-\frac{1}{2}} \exp\{2\theta^2 p_s^2/(f^4 - 2\theta\lambda_s f^2)\}.\end{aligned}$$

Now $Q^*/f^2 = \chi_k^2$ if and only if

$$\phi(\theta) = (1 - 2\theta)^{-\frac{1}{2}k}.$$

Hence

$$p_s = 0, \qquad q = 0, \tag{24}$$

and k of the λ_s must be equal to f^2 while the remaining λ_s vanish. No generality is lost in assuming

$$\lambda_1 = \lambda_2 = \cdots = \lambda_k = f^2, \qquad \lambda_{k+1} = \cdots = \lambda_{m+n} = 0. \tag{25}$$

Let $w_i = fv_i$, $i = 1, 2, \cdots, k$. From equations (21) to (25) we deduce that

$$Q^* = \sum_{s,t} q_{st}^* \zeta_s \zeta_t = \sum_{i=1}^{k} w_i^2 , \tag{26}$$

where q_{st}^* is i. of p., and the w_i are linear combinations of the ζ_s such that

$$E(w_i w_j) = f^2 \delta_{ij} . \tag{27}$$

That the conditions (26) are necessary[4] for $Q^*/f^2 = \chi_k^2$ constitutes the desired lemma.

5. Maximum Number of Degrees of Freedom. We have seen that the w_i in (26) must be of the form

$$w_i = \sum_{j=1}^{m} a_{ij} \xi_j + \sum_{j=1}^{n} b_{ij} \eta_j . \tag{28}$$

We substitute (28) and (20) into (27) and write the result in matrix form,

$$\mu AA' + \nu BB' = (a^2\mu + b^2\nu) I_k , \tag{29}$$

where I_j is the identity matrix of order j, $A^{k\times m} = (a_{ij})$, $B^{k\times n} = (b_{ij})$, and whenever a new matrix is introduced, a superscript $r \times c$ indicates that it has r rows

[4] We have incidentally proved sufficiency.

and c columns. Now if we knew that AA' and BB' were i. of p., then we could equate coefficients of μ and ν in (29) and immediately draw the desired conclusion $k \leqq m$. But that AA', BB' are i. of p. is not obvious, since this need not be true of A and B. However, we do know that the matrices

$$F = A'A, \qquad G = A'B, \qquad H = B'B$$

are i. of p. because the matrix (q_{ij}^*) of (26) is

$$\begin{pmatrix} F & G \\ G' & H \end{pmatrix}.$$

Multiplying (29) on the left by A' and on the right by A, we obtain

$$\mu F^2 + \nu GG' = (a^2\mu + b^2\nu)F. \tag{30}$$

(30) must hold identically in μ, ν. Since the coefficients of μ, ν are now i. of p., we may equate them, hence $GG' = b^2F$. Similarly multiplying (29) by B' and B, we get $G'G = a^2H$. Now[5] for any matrix M, rank M = rank $M'M$ = rank MM'. Thus rank F = rank $H = r$, say. Again, $F = A'A$, therefore r = rank $A \leqq m$. Since F is a positive[5] matrix, i. of p., there exists a non-singular $P^{m\times m}$, i. of p., such that

$$F = P'I_{m,r}P = A'A, \tag{31}$$

where $I_{j,r}$ is the $j \times j$ matrix the first r of whose diagonal elements are unity and all other elements zero. Let $T^{k\times m} = AP^{-1}$. Then

$$A = TP, \qquad T'T = I_{m,r} \tag{32}$$

from (31). Likewise we can write

$$B = U^{k\times n}R^{n\times n}, \qquad U'U = I_{n,r}, \tag{33}$$

where R is non-singular and i. of p. Then $G = A'B = P'T'UR$, hence $T'U = (P')^{-1}GR^{-1}$ is i. of p. We note

$$T = (T_1^{k\times r}, 0^{k\times(m-r)}), \qquad U = (U_1^{k\times r}, 0^{k\times(n-r)}),$$

where

$$T_1'T_1 = U_1'U_1 = I_r.$$

Since

$$T'U = \begin{pmatrix} T_1' \\ 0 \end{pmatrix}(U_1, 0) = \begin{pmatrix} T_1'U_1 & 0 \\ 0 & 0 \end{pmatrix}$$

[5] A simple proof [5b] of this useful theorem is the following: Let r = rank M, p = rank $M'M$. $p \leqq r$ since the rank of the product cannot exceed the rank of a factor. M contains r independent column vectors; the Grammian matrix of these vectors is non-singular and appears as an $r \times r$ minor in $M'M$. Hence $p \geqq r$. Furthermore, all principal minors of $M'M$ are Grammian matrices (which always have non-negative determinants), hence $M'M$ is always positive——we use this below.

is i. of p., so is its minor $V^{r\times r} = T_1'U_1$. Write

$$P = \begin{pmatrix} P_1^{r\times m} \\ P_2^{(m-r)\times m} \end{pmatrix}, \qquad R = \begin{pmatrix} R_1^{r\times n} \\ R_2^{(n-r)\times n} \end{pmatrix}.$$

Then from (32), (33),

$$A = T_1P_1, \qquad B = U_1R_1. \tag{34}$$

Substituting (34) in (29), we get

$$\mu T_1P_1P_1'T_1' + \nu U_1R_1R_1'U_1' = (a^2\mu + b^2\nu)I_k, \tag{35}$$

and multiplying by T_1' on the left, T_1 on the right,

$$\mu P_1P_1' + \nu VR_1R_1'V' = (a^2\mu + b^2\nu)I_r.$$

Again the coefficients of μ, ν are i. of p., so

$$P_1P_1' = a^2I_r,$$

$$VR_1R_1'V' = b^2I_r. \tag{36}$$

Similarly we find

$$R_1R_1' = b^2I_r. \tag{37}$$

From (36), (37), $VV' = I_r$. (35) now becomes

$$a^2\mu T_1T_1' + b^2\nu U_1U_1' = (a^2\mu + b^2\nu)I_k. \tag{38}$$

Multiplication of (38) on the right by U_1 gives

$$a^2\mu T_1V + b^2\nu U_1 = (a^2\mu + b^2\nu)U_1.$$

Hence $T_1V = U_1$, therefore $U_1U_1' = T_1T_1'$, and putting this back into (38) we have $I_k = T_1T_1'$, rank I_k = rank T_1T_1' = rank $T_1'T_1$ = rank I_r, $k = r \leqq m$.

6. Minimum Expected Length of Confidence Intervals. We now point out that of all confidence intervals (13) with $k = m$, the confidence intervals (9) have the minimum expected length. Recalling that the a_i, b_i in (19) are subject to the conditions (16), we easily find

$$\sum_{i=1}^{m} a_i^2 \geqq 1/m, \qquad \sum_{i=1}^{n} b_i^2 \geqq 1/n. \tag{39}$$

From (39) and (19) we have

$$E(l) \geqq t_{m-1,c}c_{m-1}[\mu + (m/n)\nu]^{\frac{1}{2}}/m^{\frac{1}{2}},$$

and referring to the statement at the end of section 2, the property of (9) asserted above is now obvious.

7. Asymptotic Shortness of Confidence Intervals. In conclusion we wish to compare our results with the case where the ratio of the variances, $\theta = \nu/\mu$, is known. If

$$S_x = \sum_{i=1}^{m} (x_i - \bar{x})^2, \qquad S_y = \sum_{i=1}^{n} (y_i - \bar{y})^2,$$

$$L = \bar{x} - \bar{y}, \qquad \sigma_L^2 = (\mu/m) + (\nu/n),$$

then $(L - \delta)/\sigma_L$, S_x/μ, S_y/ν are mutually independently distributed, the first normally with zero mean and unit variance, and $S_x/\mu = \chi^2_{m-1}$, $S_y/\nu = \chi^2_{n-1}$. Hence

$$(L - \delta)[(\mu m^{-1} + \nu n^{-1})(S_x\mu^{-1} + S_y\nu^{-1})/(m + n - 2)]^{-\frac{1}{2}} = t_{m+n-2}.$$

TABLE I

Values of R for $\epsilon = .95$

$m-1$ \ $n-1$	5	10	20	40	∞
5	1.15	1.20	1.23	1.25	1.28
10		1.05	1.07	1.09	1.11
20			1.03	1.03	1.05
40				1.01	1.02

TABLE II

Values of R for $\epsilon = .99$

$m-1$ \ $n-1$	5	10	20	40	∞
5	1.27	1.36	1.42	1.47	1.52
10		1.10	1.13	1.16	1.20
20			1.05	1.06	1.09
40				1.02	1.04

This relation yields the confidence intervals

$$(40) \qquad |\delta - L| \leqq t_{m+n-2,\epsilon}(m + n - 2)^{-\frac{1}{2}}(m^{-1} + \theta n^{-1})^{\frac{1}{2}}(S_x + S_y/\theta)^{\frac{1}{2}},$$

where the confidence coefficient is again ϵ. The confidence intervals (40) are known to be highly efficient; for instance they are of the shortest unbiased type [5a]. We calculate their expected length to be

$$E(l) = t_{m+n-2,\epsilon}c_{m+n-2}[\mu + (m/n)\nu]^{\frac{1}{2}}/m^{\frac{1}{2}}.$$

The ratio R of $E(l)$ for (9) to $E(l)$ for (40) is thus

$$(41) \qquad R = (t_{m-1,\epsilon}c_{m-1})/(t_{m+n-2,\epsilon}c_{m+n-2}).$$

As $k \to \infty$, $c_k \to 2$, $t_{k,\epsilon} \to t_{\infty,\epsilon}$, hence as $m \to \infty$, $R \to 1$ no matter what the values of $n \geqq m$ For small values of m the ratio of the t values in (41) is considerably >1, but this is partly offset by c_k approaching its limiting value 2 from below so that the ratio of the c's is <1. The behaviour of R for finite m is indicated in Tables I and II. Table I (II) tells us for example that with $m > 10$, and $\epsilon = .95$ (.99), the expected length of the confidence intervals (9) is at most 11 per cent (20%) longer than that of the optimum confidence intervals (40) available when the ratio θ is known. While we may conclude from $R \to 1$ as $m \to \infty$, that our solution (9) is asymptotically extremely efficient, we cannot conclude from Tables I, II that for small m (9) is inefficient, since we do not know what the lengthening effect of the extra nuisance parameter in the Behrens-Fisher problem would be on "best" confidence intervals.

8. Summary. In the terminology of the first paragraphs of sections 1 and 3 we have proved that there do not exist a linear form L and a quadratic form Q in the observations such that the quotient (10) will have the t-distribution (for *all* values of the parameters) with more than $m - 1$ degrees of freedom. We have further shown that of all confidence intervals (13) based on the t-distribution with $m - 1$ degrees of freedom, and with confidence coefficient ϵ, (9) has the minimum expected length. The quantities needed to apply our solution (9) are given by (1), (7) and (8). Finally, by comparing this solution with a known highly efficient solution for the case when the ratio of the population variances is known, it has been possible to show that at least asymptotically our confidence intervals (9) are very short.

REFERENCES

[1] W. U. Behrens, *Landwirtschaftliche Jahrbücher*, Berlin, Vol. 68 (1929), pp. 807–837.
[2] R. A. Fisher, *Annals of Eugenics*, Vol. 6 (1935), pp. 391–398; Vol. 9 (1939), pp. 174–180; Vol. 11 (1941), pp. 141–172. *Annals of Math. Stat.*, Vol. 10 (1939), pp. 383–388.
[3] J. Neyman, *Biometrika*, Vol. 32 (1941), p. 138.
[4] B. L. Welch, *Biometrika*, Vol 29 (1938), p. 360.
[5] H. Scheffé, [a] *Annals of Math. Stat*, Vol. 13 (1942), p 386. [b] *Amer. Math. Monthly*, Vol. 49 (1942), p. 101.

Further Bibliography

[6] M. S. Bartlett, *Proc. Camb. Phil. Soc*, Vol. 32 (1936), pp. 560–566. *Annals of Math. Stat*, Vol. 10 (1939) pp 129–138.
[7] P. L Hsu, *Stat. Res Mem.*, Vol. 2 (1938), pp. 1–25.
[8] H. Jeffreys, *Annals of Eugenics*, Vol. 10 (1940), pp. 48–51.
[9] E. J. G. Pitman, *Suppl J. Roy. Stat. Soc.*, Vol. 4 (1937), pp. 117–130.
[10] D M. Starkey, *Annals of Math. Stat.*, Vol. 9 (1938), pp. 201–213.
[11] P. V. Sukhatme, *Sankhyā*, Vol. 4 (1938), pp. 39–48.
[12] S S Wilks, *Annals of Math. Stat*, Vol 11 (1940), p 474.
[13] F. Yates, *Proc. Camb Phil. Soc.*, Vol. 35 (1939), pp 579–609.

AN EXTENSION OF WILKS' METHOD FOR SETTING TOLERANCE LIMITS

By Abraham Wald

Columbia University

1. Introduction. Let x be a random variable and let $f(x)$ be its probability density function. Suppose that nothing is known about $f(x)$ except that it is continuous. Let $x_1, \cdots, x_n$ be n independent observations on x. The problem of setting tolerance limits can be formulated as follows: *For some given positive values $\beta < 1$ and $\gamma < 1$ we have to construct two functions $L(x_1, \cdots, x_n)$ and $M(x_1, \cdots, x_n)$, called tolerance limits, such that the probability that*

$$\int_L^M f(t)\,dt \geq \gamma, \tag{1}$$

holds, is equal to β. This problem has recently been solved by S. S. Wilks[1] in a very satisfactory way when nothing is known about $f(x)$ except that it is continuous. Wilks proposes the following solution: Let $x_1, \cdots, x_n$ be the observed values of x arranged in order of increasing magnitude. Then $L = x_r$ and $M = x_{n-r+1}$ where r denotes a positive integer. The exact sampling distribution of the statistic $\int_{x_r}^{x_{n-r+1}} f(t)\,dt$ is derived by Wilks and this provides the solution for the problem of setting tolerance limits. A very important feature of Wilks' solution is the fact that the distribution of $\int_{x_r}^{x_{n-r+1}} f(t)\,dt$ is entirely independent of the unknown density function $f(x)$, i.e. the distribution of $\int_{x_r}^{x_{n-r+1}} f(t)\,dt$ is the same for any arbitrary continuous density function $f(x)$.

In this paper we shall give an extension of Wilks' method to the multivariate case. Let $x_1, \cdots, x_p$ be a set of p random variables with the joint probability density function $f(x_1, \cdots, x_p)$. Suppose that nothing is known about $f(x_1, \cdots, x_p)$ except that it is a continuous function of $x_1, \cdots, x_p$. A sample of n independent observations is drawn and the α-th observation on x_i is denoted by $x_{i\alpha}$ $(i = 1, \cdots, p;\ \alpha = 1, \cdots, n)$ The problem of setting tolerance limits for $x_1, \cdots, x_p$ can be formulated as follows: *For some given positive values $\beta < 1$ and $\gamma < 1$ we have to construct p pairs of functions of the observations $L_i(x_{11}, \cdots, x_{pn})$ and $M_i(x_{11}, \cdots, x_{pn})$ $(i = 1, \cdots, p)$ such that the probability that*

$$\int_{L_p}^{M_p} \cdots \int_{L_1}^{M_1} f(t_1, \cdots, t_p)\,dt_1 \cdots dt_p \geq \gamma, \tag{2}$$

[1] S. S. Wilks, "Determination of sample sizes for setting tolerance limits," *Annals of Math. Stat.*, Vol. 12 (1941).

holds, is equal to β. The functions L_i and M_i are called the lower and upper tolerance limits of x_i. A natural extension of Wilks' procedure would seem to be the following: Let $x_{i1}, \cdots, x_{in}$ be the observations on x_i arranged in order of increasing magnitude and let $L_i = x_{ir_i}$ and $M_i = x_{is_i}$, $(i = 1, \cdots, p)$ where r_i and s_i denote some integers. However, this choice of the tolerance limits does not provide a satisfactory solution of our problem, since the distribution of (2) is *not* independent of the unknown density function $f(x_1, \cdots, x_p)$. It will be shown in this paper that by a slight modification of the above procedure the distribution of (2) becomes entirely independent of the unknown density function $f(x_1, \cdots, x_p)$. In section 2 we will treat the bivariate case and in section 3 we will extend the results to multivariate distributions.

2. The bivariate case. In this section we deal with the case when $p = 2$. Let $x_{11}, \cdots, x_{1n}$ be the observations on x_1 arranged in order of increasing magnitude. We may assume that $x_{11} < x_{12} < \cdots < x_{1n}$ since the probability of an equality sign is equal to zero. We define

$$L_1 = x_{1r_1} \quad \text{and} \quad M_1 = x_{1s_1}, \tag{3}$$

where r_1 and s_1 denote some positive integers and $r_1 < s_1 \leq n$. Consider only those sample points $(x_{1\alpha}, x_{2\alpha})$ for which $x_{1r_1} < x_{1\alpha} < x_{1s_1}$, i.e. consider the sample points $(x_{1,r_1+1}, x_{2,r_1+1}), \cdots, (x_{1,s_1-1}, x_{2,s_1-1})$. Denote by $x'_{2,r_1+1}, \cdots, x'_{2,s_1-1}$ the values $x_{2,r_1+1}, \cdots, x_{2,s_1-1}$ arranged in order of increasing magnitude. We define

$$L_2 = x'_{2r_2} \quad \text{and} \quad M_2 = x'_{2s_2}, \tag{4}$$

where r_2 and s_2 denote some positive integers for which $r_2 < s_2 \leq s_1 - r_1 - 1$.

We will show that the distribution of the statistic

$$Q = \int_{L_2}^{M_2} \int_{L_1}^{M_1} f(t_1, t_2)\, dt_1\, dt_2, \tag{5}$$

is entirely independent of the unknown density function $f(x_1, x_2)$. Denote by $\varphi(x_1)$ the marginal distribution of x_1, i.e.

$$\varphi(x_1) = \int_{-\infty}^{+\infty} f(x_1, x_2)\, dx_2. \tag{6}$$

Furthermore denote by $\psi(x_2, L_1, M_1)$ the conditional distribution of x_2 calculated under the condition that $L_1 < x_1 < M_1$. Hence

$$\psi(x_2, L_1, M_1) = \frac{\int_{L_1}^{M_1} f(x_1, x_2)\, dx_1}{\int_{-\infty}^{+\infty} \int_{L_1}^{M_1} f(x_1, x_2)\, dx_1\, dx_2}. \tag{7}$$

Let

$$P = \int_{L_1}^{M_1} \varphi(t)\, dt \tag{8}$$

and

$$\bar{P} = \int_{L_2}^{M_2} \psi(t, L_1, M_1)\, dt \tag{9}$$

From (5), (8) and (9) it follows that

$$Q = P\bar{P}. \tag{10}$$

It is obvious that the distribution of P is given by Wilks' formula. Since Wilks derived the distribution only when $s_1 = n - r_1 + 1$, we will briefly give here the derivation for any integers r_1 and s_1.

Let $\int_{-\infty}^{x_{1r_1}} \varphi(t)\, dt = u$ and $\int_{x_{1s_1}}^{\infty} \varphi(t)\, dt = v$. Then the joint probability density function of u and v is given by

$$cu^{r_1-1}(1 - u - v)^{s_1-r_1-1} v^{n-s_1}\, du\, dv, \tag{11}$$

where c is a constant. We obviously have $P = 1 - u - v$. The joint density function of P and u is given by

$$cu^{r_1-1}P^{s_1-r_1-1}(1 - u - P)^{n-s_1}\, du\, dP, \tag{12}$$

where u is restricted to the interval $[0, 1 - P]$. Hence the distribution of P is given by

$$\begin{aligned} &cP^{s_1-r_1-1}\int_0^{1-P} u^{r_1-1}(1 - u - P)^{n-s_1}\, du \\ &\quad = cP^{s_1-r_1-1}(1 - P)^{n-s_1+r_1}\int_0^{1-P} \left(\frac{u}{1-P}\right)^{r_1-1}\left(1 - \frac{u}{1-P}\right)^{n-s_1} du \\ &\quad = cP^{s_1-r_1-1}(1 - P)^{n-s_1+r_1}\int_0^1 T^{r_1-1}(1 - T)^{n-s_1}\, dT \\ &\quad = c'P^{s_1-r_1-1}(1 - P)^{n-s_1+r_1}. \end{aligned}$$

Since the integral of the density function of P over the range of P must be equal to 1, we find that

$$c' = \Gamma(n + 1)/\Gamma(s_1 - r_1)\Gamma(n - s_1 + r_1 + 1).$$

Hence the probability density function of P is given by

$$\frac{\Gamma(n + 1)}{\Gamma(s_1 - r_1)\Gamma(n - s_1 + r_1 + 1)} P^{s_1-r_1-1}(1 - P)^{n-s_1+r_1}\, dP. \tag{13}$$

Since $x_{2,r_1+1}, \cdots, x_{2,s_1-1}$ can be considered as $s_1 - r_1 - 1$ independent observations on a random variable t the distribution of which is given by $\psi(t, L_1, M_1)\, dt$, for any given values L_1, M_1 the conditional distribution of $\bar{P}$ is given by the expression we obtain from (13) by substituting r_2 for r_1, s_2 for s_1 and $s_1 - r_1 - 1$ for n. Hence the conditional distribution of $\bar{P}$ is given by

$$(14) \qquad \frac{\Gamma(s_1 - r_1)}{\Gamma(s_2 - r_2)\Gamma(s_1 - r_1 - s_2 + r_2)} (\bar{P})^{s_2 - r_2 - 1}(1 - \bar{P})^{s_1 - r_1 - 1 - s_2 + r_2}\, d\bar{P}.$$

Since the expression (14) does not involve the quantities L_1 and M_1, $\bar{P}$ is distributed independently of L_1 and M_1. Hence the joint density function of P and $\bar{P}$ is given by the product of (13) and (14), i.e. by

$$(15) \qquad AP^{s_1 - r_1 - 1}(1 - P)^{n - s_1 + r_1}(\bar{P})^{s_2 - r_2 - 1}(1 - \bar{P})^{s_1 - r_1 - 1 - s_2 + r_2}\, dP\, d\bar{P},$$

where A denotes the product of the constant coefficients in (13) and (14). From (15) it follows that the joint distribution of P and $Q = P\bar{P}$ is given by

$$(16) \qquad A(1 - P)^{n - s_1 + r_1} Q^{s_2 - r_2 - 1}(P - Q)^{s_1 - r_1 - 1 - s_2 + r_2}\, dP\, dQ.$$

Since the range of P is the interval $[Q, 1]$, the distribution of Q is given by

$$(17) \qquad AQ^{s_2 - r_2 - 1} \int_Q^1 (1 - P)^{n - s_1 + r_1}(P - Q)^{s_1 - r_1 - 1 - s_2 + r_2}\, dP.$$

Let $R = P - Q$. Then we have

$$\begin{aligned} &\int_Q^1 (1 - P)^{n - s_1 + r_1}(P - Q)^{s_1 - r_1 - 1 - s_2 + r_2}\, dP \\ (18) \qquad &= \int_0^{1-Q} (1 - Q - R)^{n - s_1 + r_1} R^{s_1 - r_1 - 1 - s_2 + r_2}\, dR \\ &= (1 - Q)^{n - 1 - s_2 + r_2}(1 - Q) \int_0^1 (1 - T)^{n - s_1 + r_1} T^{s_1 - r_1 - 1 - s_2 + r_2}\, dT. \end{aligned}$$

From (17) and (18) it follows that the probability density function of Q is given by

$$(19) \qquad \frac{\Gamma(n + 1)}{\Gamma(s_2 - r_2)\Gamma(n - s_2 + r_2 + 1)} Q^{s_2 - r_2 - 1}(1 - Q)^{n - s_2 + r_2}\, dQ.$$

3. The multivariate case. We may assume that no two elements of the matrix $\| x_{i\alpha} \|$ $(i = 1, \cdots, p;\ \alpha = 1, \cdots, n)$ are equal, since the probability of this event is equal to 1. For each α let $t_\alpha (\alpha = 1, \cdots, n)$ be the point with the coordinates $x_{1\alpha}, \cdots, x_{p\alpha}$. Let $x_{11}, \cdots, x_{1n}$ be the observations on x_1 arranged in order of increasing magnitude. Then $L_1 = x_{1r_1}$ and $M_1 = x_{1s_1}$. The quantities L_i and M_i $(i = 2, \cdots, p)$ are defined in the following manner: Let S be the set of all points t_α for which

$$L_j < x_{j\alpha} < M_j \qquad (j = 1, \cdots, i - 1).$$

Arrange the i-th coordinates of the points in S in order of increasing magnitude. Then L_i is equal to the r_i-th element and M_i is equal to the s_i-th element of this ordered sequence. We will derive the distribution of

$$(20) \qquad Q_p = \int_{L_p}^{M_p} \cdots \int_{L_1}^{M_1} f(x_1, \cdots, x_p)\, dx_1 \cdots dx_p.$$

Let

$$Q_i = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \int_{L_i}^{M_i} \cdots \int_{L_1}^{M_1} f(x_1, \cdots, x_p)\, dx_1 \cdots dx_p \tag{21}$$

$$(i = 1, \cdots, p-1).$$

Denote by $\varphi_i(x_i, L_1, M_1, \cdots, L_{i-1}, M_{i-1})$ $(i = 2, \cdots, p)$ the conditional probability density function of x_i calculated under the condition that $L_j \leq x_j \leq M_j$ $(j = 1, \cdots, i-1)$. Let

$$\overline{P}_i = \int_{L_i}^{M_i} \varphi_i(x_i, L_1, M_1, \cdots, L_{i-1}, M_{i-1})\, dx_i. \tag{22}$$

We obviously have

$$Q_{i+1} = Q_i \overline{P}_{i+1} \qquad (i = 1, \cdots, p-1). \tag{23}$$

We will prove that the probability density function of Q_i is given by

$$\frac{\Gamma(n+1)}{\Gamma(s_i - r_i)\Gamma(n - s_i + r_i + 1)} Q_i^{s_i - r_i - 1}(1 - Q_i)^{n - s_i + r_i}\, dQ_i, \quad (i = 1, \cdots, p), \tag{24}$$

This is certainly true for $i = 1, 2$. We will assume that it is true for $i = j$ and we will prove it for $i = j + 1$. It is easy to see that Q_j and $\overline{P}_{j+1}$ are independently distributed and that the probability density function of $\overline{P}_{j+1}$ is given by

$$\frac{\Gamma(s_j - r_j)}{\Gamma(s_{j+1} - r_{j+1})\Gamma(s_j - r_j - s_{j+1} + r_{j+1})} \cdot (\overline{P}_{j+1})^{s_{j+1} - r_{j+1} - 1}(1 - \overline{P}_{j+1})^{s_j - r_j - 1 - s_{j+1} + r_{j+1}}\, d\overline{P}_{j+1}. \tag{25}$$

The joint distribution of Q_j and $\overline{P}_{j+1}$ is of the same form as the joint distribution of P and $\overline{P}$ in section 2. Hence the distribution of $Q_j\overline{P}_{j+1}$ can be obtained from the distribution of $Q = P\overline{P}$ by substituting r_{j+1} for r_2 and s_{j+1} for s_2. The distribution of Q is given in (19). Making the above substitution in formula (19) we obtain formula (24) for $i = j + 1$. Hence the validity of (24) is proved for $i = 1, 2, \cdots, p$. In particular, the distribution of Q_p is given by

$$\frac{\Gamma(n+1)}{\Gamma(s_p - r_p)\Gamma(n - s_p + r_p + 1)} Q_p^{s_p - r_p - 1}(1 - Q_p)^{n - s_p + r_p}\, dQ_p. \tag{26}$$

It is interesting to note that the distribution of Q_p does not depend on the integers $r_1, s_1, \cdots, r_{p-1}, s_{p-1}$. The construction of the tolerance limits L_i, M_i $(i = 1, \cdots, p)$, as proposed here, is somewhat asymmetric, since it depends on the order of the variates $x_1, \cdots, x_p$. In practical applications the asymmetry of the construction will be very slight, since in most practical cases the integers r_p and s_p will be chosen so that $(s_p - r_p - 1)/n$ will be near to 1. If, for example, $(s_p - r_p - 1)/n \geq .95$, the tolerance limits will be affected only very slightly by a permutation of the variates $x_1, \cdots, x_p$. However, it would be desirable to find a construction which is entirely independent of the order of the variates $x_1, \cdots, x_p$.

4. Tolerance regions composed of several rectangles. For the sake of simplicity we will consider here the bivariate case. All results obtained in this section can be extended without any difficulty to the multivariate case.

In section 2 the tolerance region has been a single rectangle in the plane (x_1, x_2) determined by the four lines $x_1 = L_1$, $x_1 = M_1$; $x_2 = L_2$ and $x_2 = M_2$. If the variates x_1 and x_2 are strongly correlated, a tolerance region of rectangle shape seems to be unfavorable, since it will cover an unnecessarily large area in the (x_1, x_2) plane. The situation is illustrated in figure 1 where the scatter of a

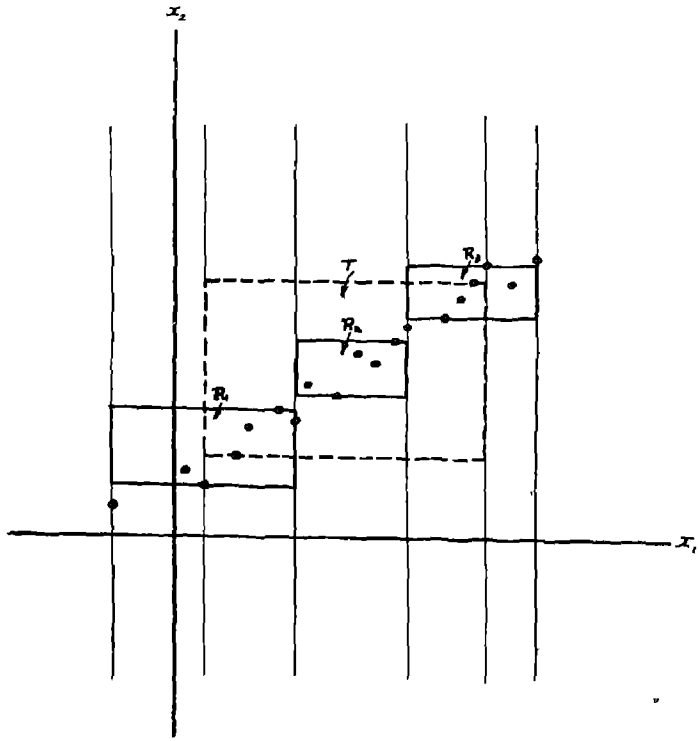


FIG. 1

bivariate sample of size $n = 19$ is shown. Suppose we choose $r_1 = 3$, $s_1 = 17$; $r_2 = 1$, $s_2 = 13$, then the tolerance region T, as defined in section 2, will be the rectangle determined by the lines $x_1 = L_1 = x_{1,3}$; $x_1 = M_1 = x_{1,17}$; $x_2 = L_2 = x'_{2,1}$; and $x_2 = M_2 = x'_{2,13}$. Now consider the tolerance region T' consisting of 3 small rectangles R_1, R_2 and R_3 defined as follows:

The rectangle R_1 is determined by the vertical lines through $x_{1,1}$ and $x_{1,7}$ and the horizontal lines through the sample points with smallest and largest ordinate, restricting ourselves to points which have abscissa values in the interior of the interval $[x_{1,1}, x_{1,7}]$. Similarly R_2 is determined by the vertical lines through

$x_{1,7}$ and $x_{1,13}$ and the horizontal lines through the sample points with largest and smallest ordinate, restricting ourselves to points with abscissa values in the interior of $[x_{1,7}, x_{1,13}]$. Finally R_3 is determined by the vertical lines through $x_{1,13}$ and $x_{1,19}$ and the horizontal lines through the sample points with largest and smallest ordinate, restricting ourselves to points whose abscissa values lie in the interior of $[x_{1,13}, x_{1,19}]$. The region T' consisting of the rectangles R_1, R_2 and R_3 has a much smaller area than the region T. As we will see later, the probability distribution of the statistic $\iint_{T'} f(x_1, x_2)dx_1\,dx_2$ is exactly the same as that of $\iint_{T} f(x_1, x_2)dx_1\,dx_2$. Thus the use of T' may be preferred to that of T.

We will consider tolerance regions T^* of the following general shape: Let $m_1, \cdots, m_k$ be k positive integers such that $1 \leq m_1, m_k \leq n$ and $m_{i+1} - m_i \geq 3$ where n is the size of the bivariate sample. Let V_i be the vertical line in the (x_1, x_2) plane given by the equation $x_1 = x_{1,m_i}$, $(i = 1, \cdots, k)$. The number of sample points which lie between the vertical lines V_i and V_{i+1} is obviously equal to $m_{i+1} - m_i - 1$. Through each point which lies between the vertical lines V_i and V_{i+1} we draw a horizontal line. In this way we obtain $m_{i+1} - m_i - 1$ horizontal lines $W_{i,1}, \cdots, W_{i,m_{i+1}-m_i-1}$ where the line $W_{i,j+1}$ is above the line $W_{i,j}$. Denote by R_{ij} $(i = 1, \cdots, k-1;\ j = 1, \cdots, m_{i+1} - m_i - 2)$ the rectangle determined by the lines V_i, V_{i+1}, $W_{i,j}$, $W_{i,j+1}$. Let T^* be a region composed of s different rectangles R_{ij}. The regions T and T' in the example illustrated in figure 1 are special cases of the type of regions T^* as described above. For the region T we have $k = 2$, $m_1 = 3$, $m_2 = 17$, $s = 12$, and for the region T' we have $k = 4$, $m_1 = 1$, $m_2 = 7$, $m_3 = 13$, $m_4 = 19$ and $s = 12$.

Let Q^* be given by $\iint_{T^*} f(x_1, x_2)dx_1\,dx_2$. We will prove that the probability density function of Q^* is given by

$$\frac{\Gamma(n+1)}{\Gamma(s)\Gamma(n-s+1)}\,Q^{*s-1}(1-Q^*)^{n-s}\,dQ^*. \tag{27}$$

Let $f_i(x_2)\,dx_2$ be the conditional distribution of x_2 under the restriction that $x_{1,m_i} < x_1 < x_{1,m_i+1}$. Thus, we have

$$f_i(x_2) = \frac{\int_{x_{1,m_i}}^{x_{1,m_i+1}} f(x_1, x_2)\,dx_1}{\int_{-\infty}^{+\infty}\int_{x_{1,m_i}}^{x_{1,m_i+1}} f(x_1, x_2)\,dx_1\,dx_2}, \tag{28}$$

Denote by $\varphi(x_1)\,dx_1$ the marginal distribution of x_1, i.e.

$$\varphi(x_1) =$$

Let

$$P_i^* = \int_{x_{1,m_i}}^{x_{1,m_i+1}} \varphi(x_1)\, dx_1 \qquad (i = 1, \cdots, k-1) \tag{29}$$

and

$$\bar{P}_i^* = \sum_j \int_{a_{ij}}^{b_{ij}} f_i(x_2)\, dx_2 \qquad (i = 1, \cdots, k-1) \tag{30}$$

where a_{ij} is the ordinate of the lower corners and b_{ij} is the ordinate of the upper corners of the rectangle R_{ij} and the summation is to be taken over all values of j for which R_{ij} is included in T^* It is clear that

$$Q^* = P_1^* \bar{P}_1^* + \cdots + P_{k-1}^* \bar{P}_{k-1}^* . \tag{31}$$

Let y be any random variable which has a continuous probability density function, say $\psi(y)\, dy$. Furthermore let $y_1, \cdots, y_n$ be n independent observations on y. Let $\psi_i(y)\, dy$ be the conditional density function of y under the condition that y is restricted to the interval $[y_{m_i}, y_{m_i+1}]$. Let

$$P' = \sum_{i,j} \int_{y_{m_i+j}}^{y_{m_i+j+1}} \psi(y)\, dy \tag{32}$$

where the summation is taken over all pairs i, j for which R_{ij} is contained in T^* Let

$$P_i' = \int_{y_{m_i}}^{y_{m_i+1}} \psi(y)\, dy, \qquad \text{and}$$

$$\bar{P}_i' = \sum_j \int_{y_{m_i+j}}^{y_{m_i+j+1}} \psi_i(y)\, dy,$$

where the summation is to be taken over all values j for which R_{ij} is contained in T^*. We obviously have

$$P' = P_1' \bar{P}_1' + \cdots + P_{k-1}' \bar{P}_{k-1}' . \tag{33}$$

It is easy to verify that (i) the joint distribution of $P_1', \cdots, P_{k-1}'$ is the same as the joint distribution of $P_1^*, \cdots, P_{k-1}^*$; (ii) the distribution of $\bar{P}_i'$ is the same as that of $\bar{P}_i^*$ $(i = 1, \cdots, k-1)$; (iii) the variates $\bar{P}_1', \cdots, \bar{P}_{k-1}'$ are independent of each other and also of $P_1', \cdots, P_{k-1}'$; (iv) the variates $\bar{P}_1^*, \cdots, \bar{P}_{k-1}^*$ are independent of each other and also of $P_1^*, \cdots, P_{k-1}^*$. Hence it follows from (31) and (33) that the distribution of Q^* is the same as that of P'. Now we will derive the distribution of P'. The expression P' can be written in the following form:

$$P' = \sum_{i=1}^{l} \int_{y_{r_i}}^{y_{s_i}} \psi(y)\, dy, \tag{34}$$

where $r_1, s_1, \cdots, r_l, s_l$ are some positive integers for which $1 \leq r_1 < s_1 < r_2 < s_2 < \cdots < r_l < s_l \leq n$. Let

$$P'' = \sum_{i=1}^{l-1} \int_{y_{r_i}}^{y_{s_i}} \psi(y)\,dy + \int_{y_{s_{l-1}}}^{y_{s_{l-1}+s_l-r_l}} \psi(y)\,dy$$

$$= \sum_{i=1}^{l-2} \int_{y_{r_i}}^{y_{s_i}} \psi(y)\,dy + \int_{y_{r_{l-1}}}^{y_{s_{l-1}+s_l-r_l}} \psi(y)\,dy. \tag{35}$$

For any fixed value $y_{s_{l-1}}$ denote by $\psi_1(y)$ the conditional probability density of y under the restriction that $y < y_{s_{l-1}}$ and by $\psi_2(y)$ the conditional distribution of y under the restriction that $y > y_{s_{l-1}}$. Let

$$P = \int_{-\infty}^{y_{s_{l-1}}} \psi(y)\,dy \qquad P_1 = \sum_{i=1}^{l-1} \int_{y_{r_i}}^{y_{s_i}} \psi_1(y)\,dy;$$

$$P_2 = \int_{y_{r_l}}^{y_{s_l}} \psi_2(y)\,dy \quad \text{and} \quad P_3 = \int_{y_{s_{l-1}}}^{y_{s_{l-1}+s_l-r_l}} \psi_2(y)\,dy.$$

Then it follows from (34) and (35) that

$$P' = PP_1 + (1 - P)P_2,$$

$$P'' = PP_1 + (1 - P)P_3. \tag{36}$$

For calculating the distributions of P_2 and P_3 we may consider the variates $y_{s_{l-1}+1}, \cdots, y_n$ as $n - s_{l-1}$ independent observations drawn from a population which has the distribution $\psi_2(y)\,dy$. Hence, the distribution of P_2 can be derived from (13) and it is easy to verify that the distribution of P_3 is the same as that of P_2. It is clear that P_2 is independent of P and P_1. Similarly P_3 is independent of P and P_1. Hence, because of (36) the distribution of P' must be the same as that of P''.

In the same way we find that the distribution of P'' is the same as the distribution of

$$P''' = \sum_{i=1}^{l-3} \int_{y_{r_i}}^{y_{s_i}} \psi(y)\,dy + \int_{y_{r_{l-2}}}^{y_{s_{l-2}+s_{l-1}+s_l-r_l-r_{l-1}}} \psi(y)\,dy$$

Thus, by induction we see that the distribution of P' is the same as the distribution of the statistic $P_0 = \int_{y_{r_1}}^{y_{r_1+s}} \psi(y)\,dy$ where $s = \sum_{i=1}^{l} (s_i - r_i)$. From (13) it follows that the distribution of P_0 is given by

$$\frac{\Gamma(n+1)}{\Gamma(s)\Gamma(n-s+1)} P_0^{s-1}(1 - P_0)^{n-s}\,dP_0.$$

Hence, we have proved that the distribution of Q^* is given by (27).

5. Summary of the results and numerical illustrations. I shall give here a summary of the results obtained and a few illustrative examples. The multivariate case being a straightforward extension of the bivariate case, I shall discuss merely the latter. Consider a pair of random variables x and y. Denote by $f(x, y)dx\,dy$ the joint probability density function of x and y and suppose that nothing is known about $f(x, y)$ except that it is continuous. A sample of n pairs of independent observations $(x_1, y_1), \cdots, (x_n, y_n)$ is drawn from this bivariate population. The sample can be represented by n points $p_1, \cdots, p_n$ in the plane (x, y), p_i being the point with the coordinates x_i and y_i. In section 2 we have dealt with the problem of finding a rectangle T in the plane (x, y), called tolerance region, such that we can state with high probability, say with probability .98 or .99, that the proportion Q of the bivariate universe included in the rectangle T is not less than a given number b, say not less than .98 or .99. The rectangle T is constructed as follows: Suppose that the points $p_1, \cdots, p_n$ are arranged in order of increasing magnitude of their abscissa values, i.e. $x_1 < x_2 < \cdots < x_n$. We draw a vertical line V_{r_1} through the point p_{r_1} and a vertical line V_{s_1} through p_{s_1} where r_1 and s_1 are positive integers such that $1 \leq r_1$, $r_1 \leq s_1 - 3$ and $s_1 \leq n$. We consider the set S consisting of the points $p_{r_1+1}, \cdots p_{s_1-1}$ which lie between the vertical lines V_{r_1} and V_{s_1}. We draw a horizontal line H_{r_2} through the point of S which has the r_2-th smallest ordinate in S. Finally a horizontal line H_{s_2} is drawn through the point of S which has the s_2-th smallest ordinate in S. The values r_2 and s_2 are positive integers for which $r_2 < s_2$. The tolerance region T is the rectangle determined by the lines V_{r_1}, V_{s_1}, H_{r_2} and H_{s_2}. The probability p that at least the porportion $b(0 < b < 1)$ of the universe is included in T is given by

$$(37) \qquad p = \int_b^1 \frac{\Gamma(n+1)}{\Gamma(s_2 - r_2)\Gamma(n - s_2 + r_2 + 1)} Q^{s_2 - r_2 - 1}(1 - Q)^{n - s_2 + r_2}\, dP.$$

It is known that if a random variable $v(0 \leq v \leq 1)$ has the distribution

$$\frac{\Gamma(c+d)}{\Gamma(c)\Gamma(d)} v^{c-1}(1 - v)^{d-1}\, dv,$$

and $2c$ and $2d$ are positive integers, then $\frac{2c}{2d}\frac{1 - v}{v}$ has the F-distribution (analysis of variance distribution) with $2d$ and $2c$ degrees of freedom. Thus,

$$(39) \qquad \frac{2(s_2 - r_2)}{2(n - s_2 + r_2 + 1)} \frac{1 - Q}{Q} = F$$

has the F-distribution with $2(n - s_2 + r_2 + 1)$ and $2(s_2 - r_2)$ degrees of freedom. From (37) it follows that p is equal to the probability that

$$F \leq \frac{2(s_2 - r_2)}{2(n - s_2 + r_2 + 1)} \frac{1 - b}{b}$$

where F has the analysis of variance distribution with $2(n - s_2 + r_2 + 1)$ and $2(s_2 - r_2)$ degrees of freedom. For the case $r_1 = 1$, $s_1 = n$, $r_2 = 1$ and $s_2 =$

$n - 2$, the following table gives the value of the sample size n which is necessary for having the probability p that at least the proportion b of the universe is included in the tolerance rectangle T.

	$b = .97$	$b = .975$	$b = .98$	$b = .985$	$b = .99$
$p = .99$	332	398	499	668	1001
$p = .95$	256	309	385	515	771

Thus, if we want the probability to be .99 that the tolerance region will include at least 98 per cent of the universe, the sample size must be 499.

In section 4 tolerance regions are considered which are composed of several rectangles. Such a tolerance region may be more favorable than a single rectangle if x and y are highly correlated. As an illustration we consider tolerance regions T^* constructed as follows: Suppose that n is divisible by 4 and the sample points $p_1, \cdots, p_n$ are arranged in order of increasing magnitude of their abscissa values. We draw the vertical lines V_0, V_1, V_2, V_3 and V_4 through the points p_1, $p_{n/4}$, $p_{n/2}$, $p_{3n/4}$ and p_n. Let $R_i (i = 1, 2, 3, 4)$ be the rectangle determined by the vertical lines V_{i-1} and V_i and the horizontal lines H_i and H_i' where H_i and H_i' are defined as follows: consider only the points which lie between the two vertical lines V_{i-1} and V_i (points on the vertical lines are excluded). From these select the point with the smallest and the point with the largest ordinate. The lines H_i and H_i' are the horizontal lines which go through these two points respectively. The tolerance region T^* is composed of the four rectangles R_1, R_2, R_3 and R_4. The number of rectangles R_i, (defined in section 4) included in T^* is equal to $s = n - 9$. Thus, according to the results of section 4 the probability distribution of the proportion Q^* of the universe included in the region T^* is given by

$$\frac{\Gamma(n+1)}{\Gamma(n-9)\Gamma(10)} (Q^*)^{n-9} (1 - Q^*)^9 \, dQ^*.$$

Numerical calculations show that for $n = 1000$ the probability is .99 that at least 98.1 per cent of the universe will be included in the tolerance region T^*.

ASYMPTOTIC FORMULAS FOR SIGNIFICANCE LEVELS OF CERTAIN DISTRIBUTIONS

By Alfred M. Peiser

Cornell University

1. Introduction. The purpose of this paper is to derive asymptotic formulas for the significance levels, or per cent points, of certain well-known statistical distributions.[1] Although we restrict ourselves here to two distributions, those of Chi-Square and of Student's t, it will be apparent that the methods used are applicable to many other distributions as well.

The following results are obtained. Let y_p be the p per cent point of the normal distribution, that is, the distribution defined by

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}v^2}\, dv, \tag{1.1}$$

so that

$$\Phi(y_p) = 1 - p. \tag{1.2}$$

If $\chi^2_{p,n}$ and $t_{p,n}$ denote the p per cent points of the Chi-Square and Student's t distributions with n degrees of freedom respectively, then

$$\chi^2_{p,n} = n + y_p\sqrt{2n} + \frac{2}{3}(y_p^2 - 1) + \frac{y_p^3 - 7y_p}{9\sqrt{2n}} + o\left(\frac{1}{\sqrt{n}}\right), \tag{1.3}$$

and

$$t_{p,n} = y_p + \frac{y_p^3 + y_p}{4n} + o\left(\frac{1}{n}\right). \tag{1.4}$$

These formulas approximate the true values of $\chi^2_{p,n}$ and $t_{p,n}$ to a high degree of accuracy. Tables of comparative values for several values of p and n are given in Section 4.

We shall need the following theorem due to Cramér [3, p. 81; see also pp. 86–87].

Theorem 1: *Let $X_1, X_2, \cdots$ be a sequence of independent, identically distributed random variables having an absolutely continuous distribution function with mean value zero, dispersion σ and finite fifth absolute moment. Let $H_n(x)$ be the distribution function of $(X_1 + \cdots + X_n)/(\sigma\sqrt{n})$, and let $n^{-\frac{1}{2}(r-2)}\lambda_r$ denote the r-th semi-invariant of $H_n(x)$. Then*

$$\Phi(x) - H_n(x) = \frac{\lambda_3}{3!\sqrt{n}}\Phi^{(3)}(x) - \frac{\lambda_4}{4!\,n}\Phi^{(4)}(x) - \frac{10\lambda_3^2}{6!\,n}\Phi^{(6)}(x) + O(n^{-3/2}). \tag{1.5}$$

[1] This problem was proposed to the author by J. H. Curtiss.

2. The Chi-Square distribution. A random variable X is said to be distributed according to Chi-Square with n degrees of freedom ($X = \chi_n^2$) if its distribution function is

$$F_n(x) = \begin{cases} \displaystyle\int_0^x \frac{t^{\frac{n}{2}-1} e^{-\frac{1}{2}t}}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}\, dt, & x \geqslant 0, \\ 0, & x < 0. \end{cases} \tag{2.1}$$

The variable $(\chi_n^2 - n)/\sqrt{2n}$ then has the distribution function

$$H_n(x) = F_n(n + x\sqrt{2n}). \tag{2.2}$$

If we write

$$\chi_{p,n}^2 = n + y_p\sqrt{2n} + a_n, \tag{2.3}$$

so that

$$F_n(\chi_{p,n}^2) = 1 - p, \tag{2.4}$$

and let $z_{pn} = y_p + a_n/\sqrt{2n}$, then $H_n(z_{pn}) = 1 - p$, and it follows from (1.1) and (1.2) that

$$\Phi(z_{pn}) - H_n(z_{pn}) = \Phi(z_{pn}) - \Phi(y_p) = \frac{1}{\sqrt{2\pi}} \frac{a_n}{\sqrt{2n}} e^{-\frac{1}{2}(y_p + \theta a_n/\sqrt{2n})^2}, \tag{2.5}$$

where $0 < \theta < 1$. Then by a theorem of Liapounoff's [3, p. 77],

$$\frac{|a_n|}{\sqrt{2n}} e^{-\frac{1}{2}(y_p + \theta a_n/\sqrt{2n})^2} \leq \frac{K \log n}{\sqrt{n}},$$

where K denotes a constant independent of n. But if $\lim_{n\to\infty} |a_n|/\sqrt{2n} = \infty$, then $\lim_{n\to\infty} H_n(z_{pn})$ is either 0 or 1. Hence $a_n = o(\sqrt{n})$.

Fisher [1, p. 81] has suggested the use of

$$\chi_{p,n}^2 \doteq \tfrac{1}{2}[y_p + \sqrt{2n - 1}]^2.$$

A closer approximation,

$$\chi_{p,n}^2 \doteq n\left[1 - \frac{2}{9n} + y_p\sqrt{\frac{2}{9n}}\right]^3,$$

has been obtained by Wilson and Hilferty [2]. It is interesting to note that, according to (1.3), this last approximation is correct to terms of the zero-th order in n.

We apply Theorem 1 to the variables $X_j = (\chi_1^2 - 1)/\sqrt{2}$, $j = 1, 2, \cdots$. Then $\sigma = 1$, and, by the additive property of the Chi-Square distribution [3, p. 45], $H_n(x)$, the distribution function of the variable $(X_1 + \cdots + X_n)/\sqrt{n}$,

is related to $F_n(x)$ by (2.2). Thus, $\lambda_3 = 2\sqrt{2}$. It follows from (1.5) and (2.3) that

$$\lim_{n\to\infty} \sqrt{2n}\,[\Phi(z_{pn}) - H_n(z_{pn})] = \lim_{n\to\infty} \frac{2}{3\sqrt{2\pi}}\,(z_{pn}^2 - 1)e^{-\frac{1}{2}z_{pn}^2}$$

$$= \frac{2}{3\sqrt{2\pi}}\,(y_p^2 - 1)e^{-\frac{1}{2}y_p^2}, \tag{2.6}$$

since $a_n = o(\sqrt{n})$. Then by (2.5) and (2.6)

$$\lim_{n\to\infty} a_n = \tfrac{2}{3}(y_p^2 - 1).$$

According to (2.3) we may now write

$$\chi_{p,2n}^2 = 2n + 2y_p\sqrt{n} + 2r_p + 2b_n\,,$$

where

$$r_p = \tfrac{1}{3}(y_p^2 - 1), \tag{2.7}$$

and $b_n = o(1)$. A simple change of variables in (2.1) yields

$$F_{2n}(\chi_{p,2n}^2) = \int_{\frac{n+r_p}{\sqrt{n}}}^{y_p+\frac{b_n}{\sqrt{n}}} \frac{\sqrt{n}e^{-n}n^n}{\Gamma(n+1)}\,e^{-v\sqrt{n}-r_p}\left[1 + \frac{v}{\sqrt{n}} + \frac{r_p}{n}\right]^{n-1} dv. \tag{2.8}$$

If we let

$$J_n = \int_{y_p}^{y_p+\frac{b_n}{\sqrt{n}}} \frac{1}{\sqrt{2\pi}}\,e^{-\frac{1}{2}v^2}\,dv, \tag{2.9}$$

then

$$nJ_n = \frac{b_n\sqrt{n}}{\sqrt{2\pi}}\,e^{-\frac{1}{2}(y_p+\delta_n)^2}, \tag{2.10}$$

where $\delta_n = o(1)$. By (1.2) and (2.4),

$$J_n = \Phi\left(y_p + \frac{b_n}{\sqrt{n}}\right) - F_{2n}(\chi_{p,2n}^2). \tag{2.11}$$

Using Stirling's formula for $\Gamma(n+1)$ in (2.8), (2.11) becomes

$$J_n = \int_{y_p+\frac{b_n}{\sqrt{n}}}^{\infty} \frac{e^{-\frac{1}{2}v^2}}{\sqrt{2\pi}}\left[\exp\left\{-\frac{1}{12n} + O(n^{-3})\right.\right.$$

$$\left.\left. + (n-1)\log\left(1 + \frac{v}{\sqrt{n}} + \frac{r_p}{n}\right) + \frac{v^2}{2} - v\sqrt{n} - r_p\right\} - 1\right] dv$$

$$= \int_{y+\frac{b_n}{\sqrt{n}}}^{\infty} \frac{e^{-\frac{1}{2}v^2}}{\sqrt{2\pi}}\,[e^{A_n} - 1]\,dv,$$

where $A_n = \frac{1}{n}\left(-\frac{1}{12} - r_p - \frac{r_p^2}{2} + \frac{v^2}{2} + v^2 r_p - \frac{v^4}{4}\right) + \frac{1}{\sqrt{n}}\left(\frac{v^3}{3} - v - vr_p\right) + f_n(v)$, $nf_n(v)$ being dominated by $P(|v|)$, where P is a polynomial in v independent of n, and $f_n(v) = O(n^{-3/2})$. Then

$$J_n = \int_{y_p + \frac{b_n}{\sqrt{n}}}^{\infty} \frac{e^{-\frac{1}{2}v^2}}{n\sqrt{2\pi}}\left[-\frac{1}{12} - r_p - \frac{r_p^2}{2} + v^2\left(1 + 2r_p + \frac{r_p^2}{2}\right)\right.$$

$$\left. - v^4\left(\frac{7}{12} + \frac{r_p}{3}\right) + \frac{v^6}{18}\right] dv + \int_{y_p + \frac{b_n}{\sqrt{n}}}^{\infty} \frac{e^{-\frac{1}{2}v^2}}{\sqrt{2\pi}n}\left[\frac{v^3}{3} - v - vr_p\right] dv \tag{2.12}$$

$$+ \int_{y_p + \frac{b_n}{\sqrt{n}}}^{\infty} \frac{e^{-\frac{1}{2}v^2}}{\sqrt{2\pi}} g_n(v)\, dv + \int_{y_p + \frac{b_n}{\sqrt{n}}}^{\infty} \frac{e^{-\frac{1}{2}v^2}}{\sqrt{2\pi}}\left(\sum_{j=3}^{\infty} \frac{A_n^j}{j!}\right) dv,$$

where $g_n(v)$ has the same properties given above for $f_n(v)$. If we call these last integrals K_1, K_2, K_3 and K_4 respectively, we see that

$$\lim_{n\to\infty} nK_3 = \int_{y_p}^{\infty} \frac{e^{-\frac{1}{2}v^2}}{\sqrt{2\pi}} \lim_{n\to\infty} ng_n(v)\, dv = 0. \tag{2.13}$$

Also, since A_n^j, $j = 1, 2, \cdots$ is dominated by $P_j(|v|)$, $P_j(v)$ being a polynomial in v independent of n, we see that

$$\sum_{j=3}^{\infty} \int_{y_p + \frac{b_n}{\sqrt{n}}}^{\infty} \frac{e^{-\frac{1}{2}v^2}}{\sqrt{2\pi}} \frac{|A_n|^j}{j!}\, dv \leq \sum_{j=3}^{\infty} e^{-\frac{1}{2}\left(y_p + \frac{b_n}{\sqrt{n}}\right)^2} \frac{Q_j\left(y_p + \frac{b_n}{\sqrt{n}}\right)}{j!},$$

where Q_j is a polynomial. Since this last sum converges, we have

$$nK_4 = \sum_{j=3}^{\infty} \int_{y_p + \frac{b_n}{\sqrt{n}}}^{\infty} \frac{e^{-\frac{1}{2}v^2}}{\sqrt{2\pi}} \frac{nA_n^j}{j!}\, dv, \tag{2.14}$$

and by the uniform convergence of (2.14),

$$\lim_{n\to\infty} nK_4 = \sum_{j=3}^{\infty} \int_{y_p}^{\infty} \frac{e^{-\frac{1}{2}v^2}}{j!\sqrt{2\pi}} \lim_{n\to\infty} nA_n^j\, dv = 0, \tag{2.15}$$

since $A_n^j = O(n^{-j/2})$.

Integrating by parts we obtain

$$nK_2 = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(y_p + \frac{b_n}{\sqrt{n}}\right)^2}\left(\frac{2}{3} y_p b_n + \frac{b_n^2}{3\sqrt{n}}\right),$$

and since $b_n = o(1)$,

$$\lim_{n\to\infty} nK_2 = 0. \tag{2.16}$$

Further integration by parts and the use of (2.7) yields

$$\lim_{n\to\infty} nK_1 = \frac{e^{-\frac{1}{2}y_p^2}}{36\sqrt{2\pi}}(y_p^3 - 7y_p). \tag{2.17}$$

Then, by (2.10), (2.12), (2.13), (2.15), (2.16) and (2.17),

$$\lim_{n\to\infty} b_n\sqrt{n} = \frac{1}{36}(y_p^3 - 7y_p),$$

so that

$$\chi^2_{p,2n} = 2n + 2y_p\sqrt{n} + \frac{2}{3}(y_p^2 - 1) + \frac{y_p^3 - 7y_p}{18\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right).$$

Equation (1.3) now follows at once.

3. Student's t. If the random variable Y_n has the distribution function $\Phi(x/\sqrt{n})$, then $t_n = Y_n/\chi_n$ is distributed according to Student's distribution for n degrees of freedom and has the distribution function

$$G_n(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{n\pi}} \frac{\Gamma[\frac{1}{2}(n+1)]}{\Gamma[\frac{1}{2}n]} \left(1 + \frac{t^2}{n}\right)^{-\frac{1}{2}(n+1)} dt.$$

If $\sigma = \sqrt{n/(n-2)}$, the variable t_n/σ then has the distribution function

$$H_n(x) = G_n(x\sigma). \tag{3.1}$$

If we write

$$t_{p,n} = y_p + a_n, \tag{3.2}$$

so that

$$G_n(t_{p,n}) = 1 - p, \tag{3.3}$$

and let $z_{pn} = t_{p,n}/\sigma$, then $H_n(z_{pn}) = 1 - p$, and it follows from (1.1) and (1.2) that

$$\begin{aligned}\Phi(z_{pn}) - H_n(z_{pn}) &= \Phi(z_{pn}) - \Phi(y_p) \\ &= \frac{1}{\sqrt{2\pi}}\left[y_p\left(\frac{1}{\sigma} - 1\right) + \frac{a_n}{\sigma}\right] e^{-\frac{1}{2}\left[y_p + \theta\left(y_p\left(\frac{1}{\sigma}-1\right)+a_n\right)\right]^2},\end{aligned} \tag{3.4}$$

where $0 < \theta < 1$. Then by Liapounoff's Theorem [3, p. 77],

$$\left|y_p\left(\frac{1}{\sigma} - 1\right) + \frac{a_n}{\sigma}\right| e^{-\frac{1}{2}\left[y_p + \theta\left(y_p\left(\frac{1}{\sigma}-1\right)+a_n\right)\right]^2} \leqslant \frac{K \log n}{\sqrt{n}},$$

where K denotes a constant independent of n. But if $\lim_{n\to\infty} |a_n| = \infty$, then $\lim_{n\to\infty} H_n(z_{pn})$ is either 0 or 1.

Hence $a_n = o(1)$.

We apply Theorem 1 to the variables $X_j = Y_n/\chi_n$, $j = 1, 2, \cdots$. Then $\sigma = \sqrt{n}/(n-2)$, and by the additive property of the normal distribution, $H_n(x)$, the distribution function of $(X_1 + \cdots + X_n)/(\sigma\sqrt{n})$, satisfies the relation (3.1). Thus $\lambda_3 = 0$ and $\lambda_4 = 6n/(n-4)$. It follows from (1.5) and (3.2) that

$$\lim_{n\to\infty} n[\Phi(z_{pn}) - H_n(z_{pn})] = \lim_{n\to\infty} \frac{n}{4(n-4)\sqrt{2\pi}} (z_{pn}^3 - 3z_{pn})e^{-\frac{1}{2}z_{pn}^2} \tag{3.5}$$
$$= \frac{1}{4\sqrt{2\pi}} (y_p^3 - 3y_p)e^{-\frac{1}{2}y_p^2},$$

since $a_n = o(1)$. By (3.4) and (3.5) we have

$$\lim_{n\to\infty} n\left[y_p\left(\frac{1}{\sigma} - 1\right) + \frac{a_n}{\sigma}\right] = \frac{y_p^3 - 3y_p}{4}.$$

But $\lim_{n\to\infty} n(1-\sigma)/\sigma = -1$, so that

$$\lim_{n\to\infty} na_n = \frac{y_p^3 + y_p}{4}.$$

Hence

$$a_n = \frac{y_p^3 + y_p}{4n} + o\left(\frac{1}{n}\right),$$

and equation (1.4) follows at once.

4. Tables. The following tables compare the true values of $\chi_{p,n}^2$ and $t_{p,n}$ with those obtained from (1.3) and (1.4). The true values [4], [5], (to three decimal places) are shown in *italics*.

TABLE 1

$\chi_{p,n}^2$

n \ p	.01	.05	.1	.5	.9
10	23.253 *23.209*	18.318 *18.307*	15.989 *15.987*	9.333 *9.342*	4.875 *4.865*
30	50.908 *50.892*	43.777 *43.773*	40.257 *40.256*	29.333 *29.336*	20.600 *20.599*
50	76.163 *76.154*	67.507 *67.505*	63.168 *63.167*	49.333 *49.335*	37.689 *37.689*
100	135.811 *135.807*	124.343 *124.342*	118.499 *118.498*	99.333 *99.334*	82.358 *82.358*

TABLE II

$t_{p,n}$

n \ p	.0125	.025	.05	.125	25
10	2.579 *2.634*	2.197 *2.228*	1.797 *1.813*	1.212 *1.221*	0 700 *0.700*
30	2.354 *2.360*	2.039 *2.042*	1.696 *1.697*	1.171 *1.173*	0.683 *0.683*
60	2.298 *2 299*	2.000 *2.000*	1.670 *1.671*	1.161 *1.162*	0.679 *0.679*
120	2.270 *2.270*	1.980 *1.980*	1.658 *1 658*	1.156 *1.156*	0.677 *0.677*

REFERENCES

[1] R. A. Fisher, *Statistical methods for research workers*, Oliver and Boyd, 1925.

[2] E B Wilson and M. M. Hilferty, *National Academy of Sciences Proc*, Vol. 17 (1931), p. 687.

[3] H Cramér, *Random variables and probability distributions*, Cambridge University Press, 1937.

[4] M. Merrington, "Tables of percentage points of the t distribution," *Biometrika*, Vol. 32 (April 1942), p 300

[5] M Merrington, "Numerical approximation to the percentage points of the χ^2 distribution," *Biometrika*, Vol. 32 (October 1941), p. 200.

GENERALIZATION OF POINCARÉ'S FORMULA IN THE THEORY OF PROBABILITY

By Kai Lai Chung

Tsing Hua University, Kunming, China

Let $p_{[m]}(1, \cdots, n)$, $(0 \leq m \leq n)$ denote the probability of the occurrence of exactly m events among the n arbitrary events $E_1, \cdots, E_n$; and $p_m(1, \cdots, n)$ $(1 \leq m \leq n)$ that of at least m. Let $p_{\nu_1 \cdots \nu_i}$ $(1 \leq i \leq n)$, where $(\nu_1 \cdots \nu_i)$ is a combination (without repetition) out of $(1, \cdots, n)$, denote the probability of the occurrence of $E_{\nu_1}, \cdots, E_{\nu_i}$ (without regard to the other events); and

$$S_0 = 1, \qquad S_i = \sum_{(\nu_1 \cdots \nu_i)} p_{\nu_1 \cdots \nu_i},$$

where the summation extends to all the combinations with i members out of $(1, \cdots, n)$.

Then Poincaré's formula may be written as follows:

$$p_{[0]}(1, \cdots, n) = \sum_{i=0}^{n} (-1)^i S_i.$$

An equivalent formula is:

$$p_1(1, \cdots, n) = \sum_{i=1}^{n} (-1)^{i-1} S_i.$$

The following conventions concerning the binomial coefficients are made:

$$\binom{0}{0} = 1, \qquad \binom{a}{b} = 0 \qquad \text{if } a < b \text{ or } b < 0.$$

Two generalizations, possibly due to de Mises, are

$$p_{[m]}(1, \cdots, n) = \sum_{i=m}^{n} (-1)^{(i-m)} \binom{i}{m} S_i;$$

$$p_m(1, \cdots, n) = \sum_{i=m}^{n} (-1)^{(i-m)} \binom{i-1}{m-1} S_i.$$

We notice that the probabilities appearing on the left-hand sides of these formulas are symmetrical with respect to the set of suffixes $(1, \cdots, n)$, and the sums on the right-hand sides are symmetrical in the same way.

As a natural generalization let us consider a probability which is symmetrical with respect to certain sub-sets of $(1, \cdots, n)$. We divide the n events into r sets:

$$E_{\nu_{11}}, \cdots, E_{\nu_{1n_1}}; E_{\nu_{21}}, \cdots, E_{\nu_{2n_2}}; \cdots; E_{\nu_{r1}}, \cdots, E_{\nu_{rn_r}};$$

where $n_1 + n_2 + \cdots + n_r = n$. And we ask for the probability that out of the first set of n_1 events exactly m_1 events occur; and out of the second set of n_2 events exactly m_2 events occur; and so on; and finally, out of the rth set of n_r events exactly m_r events occur. When this problem is solved the analogous problem

in which we replace some of the words "exactly" by "at least" can also be solved.

We denote the required probability by the left-hand side of the following *generalized Poincaré's formula*:

$$p_{[m_1],[m_2],\cdots,[m_r]}(\nu_{11}\cdots\nu_{1n_1};\nu_{21}\cdots\nu_{2n_2};\cdots;\nu_{r1}\cdots\nu_{rn_r})$$

$$= \sum_{i_1=m_1}^{n_1}\sum_{i_2=m_2}^{n_2}\cdots\sum_{i_r=m_r}^{n_r}(-1)^{i_1+i_2+\cdots+i_r-m_1-m_2-\cdots-m_r}\binom{i_1}{m_1}\binom{i_2}{m_2}\cdots\binom{i_r}{m_r}S_{i_1,i_2,\cdots,i_r}, \tag{1}$$

where

$$S_{i_1,i_2,\cdots,i_r} = \Sigma p_{\alpha_{11}\cdots\alpha_{1i_1}\alpha_{21}\cdots\alpha_{2i_2}\cdots\alpha_{r1}\cdots\alpha_{ri_r}},$$

the summation extending to all those combinations of α's such that for every $k = 1, \cdots, r$, $(\alpha_{k1}\cdots\alpha_{ki_k})$ is a combination of i_k members out of $(\nu_{k1}\cdots\nu_{kn_k})$.

PROOF: Let $p_{[\nu_1\cdots\nu_s]}$ denote the probability of the occurrence of the events $E_{\nu_1}, \cdots, E_{\nu_s}$ and *these only* out of $E_1, \cdots, E_n$. It is well-known and also easily seen that

$$p_{\alpha_1\cdots\alpha_a} = \sum_{b=0}^{n-a}\sum_{\beta} p_{[\alpha_1\cdots\alpha_a\beta_1\cdots\beta_b]}$$

where for a fixed b the second summation extends to all the combinations $(\beta_1\cdots\beta_b)$ of b members out of the "difference set" $(1, \cdots, n) - (\alpha_1\cdots\alpha_a)$.

Now let each p in each S on the right-hand side of (1) be decomposed into a sum of the $p_{[\nu_1\cdots\nu_s]}$'s in the last-written way. Consider a fixed

$$p_{[\mu_{11}\cdots\mu_{1j_1}\mu_{21}\cdots\mu_{2j_2}\cdots\mu_{r1}\cdots\mu_{rj_r}]},$$

where for every $k = 1, \cdots, r$, $(\mu_{k1}, \cdots, \mu_{kj_k})$ is a combination of j_k members out of $(\nu_{k1}\cdots\nu_{kn_k})$. It appears once in exactly $\binom{j_1}{i_1}\binom{j_2}{i_2}\cdots\binom{j_r}{i_r}$ terms in $S_{i_1,i_2,\cdots,i_r}$. Hence, its total contribution to the right-hand side of (1) is

$$\sum_{i_1=m_1}^{n_1}\sum_{i_2=m_2}^{n_2}\cdots\sum_{i_r=m_r}^{n_r}(-1)^{i_1+i_2+\cdots+i_r-m_1-m_2-\cdots-m_r}\cdot\binom{i_1}{m_1}\binom{i_2}{m_2}\cdots\binom{i_r}{m_r}\binom{j_1}{i_1}\binom{j_2}{i_2}\cdots\binom{j_r}{i_r}$$

$$= \prod_{k=1}^{r}\binom{m_k}{j_k}\sum_{i_k=m_k}^{n_k}(-1)^{i_k-m_k}\binom{j_k-m_k}{i_k-m_k}$$

$$= \prod_{k=1}^{r}\binom{j_k}{m_k}\sum_{h_k=0}^{j_k-m_k}(-1)^{h_k}\binom{j_k-m_k}{h_k}$$

$$= \prod_{k=1}^{r}\binom{j_k}{m_k}\cdot\begin{cases}1 & \text{if } j_k = m_k\\ 0 & \text{if otherwise}\end{cases}$$

$$= \begin{cases}1 & \text{if } j_k = m_k \text{ for every } k = 1, \cdots, r\\ 0 & \text{otherwise.}\end{cases}$$

Therefore after the decompositions and the collecting of terms, the only p's remaining on the right-hand side of (1) are those in which for every $k = 1, \cdots, j$ we have $j_k = m_k$. Thus the right-hand side is reduced to

$$\Sigma p_{(\mu_{11}\cdots\mu_{1m_1}\mu_{21}\cdots\mu_{2m_2}\cdots\mu_{r1}\cdots\mu_{rm_r})},$$

where the summation extends to all those combinations of μ's such that for every $k = 1, \cdots, r$, $(\mu_{k1} \cdots \mu_{km_k})$ is a combination of m_k members out of $(\nu_{k1} \cdots \nu_{kn_k})$. This is clearly equal to the left-hand side of (1). Q. E. D.

If we replace "exactly m_k" by "at least m_k" in the definition of the probability just considered, we replace in our notation the square-bracketed $[m_k]$ by an unbracketed m_k and we replace in our formula $\binom{i_k}{m_k}$ by $\binom{i_k - 1}{m_k - 1}$. This is proved as before, noting that we have

$$\sum_{i_k=m_k}^{n_k} (-1)^{i_k-m_k}\binom{i_k - 1}{m_k - 1}\binom{j_k}{i_k} = 1 \qquad \text{for } j_k = m_k, \cdots, n_k;$$

and identity which can be proved by induction on j_k.

A parallel generalization of Poincaré's formula is as follows: We ask for the probability that *either* out of the first set exactly m_1 events occur; *or* out of the second exactly m_2 ; $\cdots$; *or* finally, out of the rth set exactly m_r. That is, instead of repeated conjunctions we may consider repeated disjunctions. We denote the required probability by the left hand side of (2), then it is given in terms of the p's defined above in (1) by the right-hand side below:

$$(2)\quad W_{[m_1],[m_2],\cdots,[m_r]}(\nu_{11}, \cdots, \nu_{1n_1}; \nu_{21} \cdots \nu_{2n_2}; \cdots; \nu_{r1} \cdots \nu_{rn_r}) = p_{m_1,m_2,\cdots,m_r} - p_{m_1+1,m_2+1,\cdots,m_r+1}.$$

Other events symmetrical with respect to each of the sub-sets, in whose definition the words "and", "or", "exactly", "at least" appear arbitrarily, may be considered.

Lastly, we only mention that as a first application the formula (1) can be used to establish the formula

$$(n - k)\Sigma p_m(\nu_1 \cdots \nu_k) = (k + 1 - m)\Sigma p_m(\nu_1 \cdots \nu_{k+1}) + m\Sigma p_{m+1}(\nu_1 \cdots \nu_{k+1}),$$

first obtained by P. L. Hsu. For its significance we may refer to [1], and a continuation of that paper to be published shortly.

REFERENCE

[1] K. L. Chung, "On the probability of the occurrence of least m events among n arbitrary events," *Annals of Math. Stat.*, Vol. 12 (September 1941).

TABLES FOR TESTING RANDOMNESS OF GROUPING IN A SEQUENCE OF ALTERNATIVES

By Frieda S. Swed and C. Eisenhart

University of Wisconsin

When two different kinds of objects are arranged along a line they will form two or more distinct groups of like objects. Thus, in the arrangement: *aabbbab*, there are 3 a's and 4 b's forming 4 groups. In general, if there are m objects of one kind and n objects of another kind, there are in all

$$C_m^{m+n} = C_n^{m+n}$$

different arrangements possible. There will be no loss of generality if we assume that $m \leq n$

If u is defined to be the number of distinct groups of like objects in any one arrangement, then the proportion of arrangements yielding u' or less groups is[1]

$$P\{u \leq u'\} = \frac{1}{C_m^{m+n}} \sum_{u=2}^{u'} f_u, \tag{1}$$

where

$$f_u = 2C_{k-1}^{m-1} \cdot C_{k-1}^{n-1}, \quad \text{when} \quad u = 2k, \quad \text{i.e. } u \text{ is even},$$

and

$$f_u = C_{k-1}^{m-1}\ C_{k-2}^{n-1} + C_{k-2}^{m-1} \cdot C_{k-1}^{n-1}, \quad \text{when} \quad u = 2k - 1, \quad \text{i.e. } u \text{ is odd},$$

for $k = 1, 2, \cdots, m + 1$. For example, if $m = n = 5$, then

$$P\{u = 2\} = \frac{f_2}{C_5^{10}} = \frac{2\{C_0^4 \cdot C_0^4\}}{C_5^{10}} = \frac{1}{126},$$

$$P\{u = 3\} = \frac{f_3}{C_5^{10}} = \frac{C_1^4\ C_0^4 + C_0^4\ C_1^4}{C_5^{10}} = \frac{8}{252}.$$

In a random arrangement (1) is the probability of $u \leq u'$.

The following tables have been prepared for use in testing data for randomness and for testing whether two samples are from the same population. Table I gives $P\{u \leq u'\}$ to 7 decimal places for $m \leq n \leq 20$ with a range of m from 2 to 20 inclusive whereas Table II gives correct values for u_ϵ for $\epsilon = .005, .01, .025, .05, 95, .975, .99$ and 995, where u_ϵ is the *largest* integer, u', for which $P\{u \leq u'\} \leq \epsilon$ when $\epsilon < .50$, and is the *smallest* integer, u', for which $P\{u \leq u'\} \geq \epsilon$ when $\epsilon > 50$ This table was obtained from Table I and covers the same

[1] W L Stevens, "Distribution of Groups in a Sequence of Alternatives" (*Annals of Eugenics*, Vol. IX, Part I (1939) pp 10–17)

A Wald and J Wolfowitz, "On a Test Whether Two Samples are from the Same Population" (*Annals of Math Stat*, Vol XI, No 2, June (1940) pp 147–162).

range of values of m and n. Table III gives values of u_ϵ for $m = n$ from 10 to 100. These values of u_ϵ were obtained by using the normal approximation given on page 151 of the Wald-Wolfowitz paper together with a correction for continuity not given in their article—this correction improved the approximation for small values of m and n. The values of u_ϵ for $m = n = 10$ through 20 are included in Table III although they can be obtained from Table II in order to check on the adequacy of the approximation. These values obtained with the approximation check with those of Table II except for the five underscored values. It appears that the approximation will be adequate in general for $m = n \geq 20$.

To illustrate the use of these tables to test randomness of an arrangement,[2] consider a case where one might suspect nonrandomness and, more specifically, expect too few groups. The arrangement of diseased and healthy plants in a row of a field might be such a case. For example, we might have the following plant arrangement:

$$H\ H\ H\ H\ H\ H\ H\ H\ H\ H\ H\ D\ H\ D\ D\ D\ D\ H\ H\ H\ H\ H\ H\ H\ H\ H,$$

where

$m = 5$, the number of diseased plants present,
$n = 20$, the number of healthy plants present,
$u' = 5$, the number of groups actually formed.

From Table I the probability associated with this arrangement is found to be .018,3512, which is the probability of $u \leq u'$. Since $P < .05$, we might elect to regard this as evidence of a tendency for the disease to be nonrandomly distributed among the plants in a row, knowing that if we look for an explanation of "clustering" whenever $P\{u \leq u'\} \leq .05$ we may expect to follow a false scent not more than one time in twenty in the long run.

When a control chart[3] suggests the presence of assignable causes of variation in a manufactured product flowing from a production line, an examination of various types of runs, e.g. the lengths and relative frequency of runs above and below the median of a sequence of values, may assist in diagnosing the nature of the cause. Dr. Walter A. Shewhart has given us such an instance: A sequence of observations dealing with corrosion suggested the presence of an assignable cause of variation. By the use of run charts an assignable cause of variation was tracked down in the measuring apparatus and an attempt was made to eliminate it. The original sequence examined with regard to runs above and below the median of the sequence exhibited an unexpectedly large number of runs of length 7 or more and as a result a significantly low value of

[2] W. L. Stevens (ibid).

[3] American Defense Emergency Standards Z1.1 and Z1.2 entitled "Guide for Quality Control" and "Control Chart Method of Analyzing Data" and American War Standard Z1.3 entitled "Control Chart Method of Controlling Quality During Production" (published by the American Standards Association, New York City)

u, and, if the assignable cause were not completely eliminated in the new design, we might expect too large a proportion of long runs above and below the median, and, hence, too few total runs. A sequence of 40 observations taken with the new measuring device yielded a total of 15 runs above and below the median of the sequence which is significantly fewer than would be expected to arise under a state of statistical control, since for $m = n = 20$, $P\{u \leq 15\} = .038$. This sequence is of special interest since the occurrence of too few runs suggested the assignable cause had not been entirely eliminated although no especially long runs, say of length 7 or more, occurred in this sequence, so that from the point of view of length of runs without regard to their number the assignable cause might have been judged to have been eliminated.

As an instance where too many groups would be the probable alternative to randomness consider the arrangement of occupied and unoccupied seats at a lunch counter about half an hour before the popular lunch hour begins. In such a case the critical region would be $u \geq u'$ and the appropriate probability would be $P = 1 - P\{u \leq u' - 1\}$. Such a situation was observed and yielded the following arrangement of empty and occupied seats along the lunch counter:

$$E\,O\,E\,E\,O\,E\,E\,E\,O\,E\,E\,E\,O\,E\,O\,E,$$

$$\begin{aligned} m &= 5, \\ n &= 11, \\ u' &= 11, \\ P &= 1 - .942{,}3077 = .057{,}6923; \end{aligned}$$

and though this probability is not quite significant, the arrangement observed has the maximum number of groups of empty and occupied seats for the m and n of the size observed since no two occupied seats are adjacent. However, if another customer had entered and sat either in the 5th empty seat from the left or in the 8th empty seat, the number of groups would have been increased by two and the situation would be:

$$\begin{aligned} m &= 6, \\ n &= 10, \\ u' &= 13, \\ P &= 1 - \ 989{,}5105 = .010{,}4895. \end{aligned}$$

This P value is significant, and for this assumed case, as well as for the actual case observed, the arrangement of E's and O's has the maximum number of groups of like objects. Certainly both of these cases exhibit too many groups to be considered random arrangements.

The use of these tables to test whether two samples constitute independent random samples from the same population[4] can be illustrated by using the data of Snedecor's Example 4.11 on page 75 of his *Statistical Methods* (3d edition)

[4] A. Wald and J. Wolfowitz (ibid) have pointed out that exceptionally small values of u' are to be regarded as evidence for rejecting this null hypothesis.

which gives daily gains in two lots of steer calves on two different rations. The daily rates of gain given for the two lots are:

I. 1.95, 2.17, 2.06, 2.11, 2.24, 2.52, 2.04, 1.95;
V. 1.82, 1.85, 1.87, 1.74, 2.04, 1.78, 1.76, 1.86.

Arranging these rates in order of magnitude, designating a calf on ration I by italics and one from V by ( ), we have

(1.74), (1.76), (1.78), (1.82), (1.85), (1.86), (1.87), *1.95*, *1.95*, (2.04), *2.04*, *2.06*, *2.11*, *2.17*, *2.24*, *2.52*.

Whence

$$m = 8,$$

$$n = 8,$$

$$u' = 4,$$

$$P = .008,8578.$$

Accordingly, at either the .05 or .01 level of significance rejection of the null hypothesis that the two samples constitute independent random samples from the same population is indicated.

For these data we note the fact that having two identical values, i.e. 2.04, in the two lots did not alter the number of groups regardless of whether they were recorded as (2.04), *2.04* or as *2.04*, (2.04). However, such duplications in general may be more bothersome, since they may yield different values of u' depending on the order in which they are considered. In such instances both possible orders should be considered.

The merit of this test is that it employs a minimum of assumptions—merely that the common population be continuous, and that the samples be drawn at random independently. Its principal defect is its lack of power. As a consequence gross disparity between the samples is generally required to render $u' \leq u_\epsilon$. Therefore, when additional assumptions are tenable, tests utilizing them should be employed.

Most of the computing and checking of these tables was done by Frieda S. Swed, Philip Ritz and Beatrice E. Kelley with some assistance from Jay Grodman, Edward Halamka and Mrs. Henry Wallman. Also, Duane Borst and Francis Cox helped with the typing and the proofing of the tables.

TABLE I

$P\{u \leq u'\}$

When m = n, the largest possible value of u' is 2m, when m < n, the largest possible value of u' is 2m + 1.

n = 2

n \ u'	2	3	4
2	.333,3333	.666,6667	1.
3	.200,0000	.500,0000	.900,0000
4	.133,3333	.400,0000	.800,0000
5	.095,2381	.333,3333	.714,2857
6	.071,4286	.285,7143	.642,8571
7	.055,5556	.250,0000	.583,3333
8	.044,4444	.222,2222	.533,3333
9	.036,3636	.200,0000	.490,9091
10	.030,3030	.181,8182	.454,5455
11	.025,6410	.166,6667	.423,0769
12	.021,9780	.153,8462	.395,6044
13	.019,0476	.142,8571	.371,4286
14	.016,6667	.133,3333	.350,0000
15	.014,7059	.125,0000	.330,8824
16	.013,0719	.117,6471	.313,7255
17	.011,6959	.111,1111	.298,2456
18	.010,5263	.105,2632	.284,2105
19	.009,5238	.100,0000	.271,4286
20	.008,6580	.095,2381	.259,7403

n = 3

n \ u'	2	3	4	5	6
3	.100,0000	.300,0000	.700,0000	.900,0000	1.
4	.057,1429	.200,0000	.542,8571	.800,0000	.971,4286
5	.035,7143	.142,8571	.428,5714	.714,2857	.928,5714
6	.023,8095	.107,1429	.345,2381	.642,8571	.880,9524
7	.016,6667	.083,3333	.283,3333	.583,3333	.833,3333
8	.012,1212	.066,6667	.236,3636	.533,3333	.787,8788
9	.009,0909	.054,5455	.200,0000	.490,9091	.745,4545
10	.006,9930	.045,4545	.171,3287	.454,5455	.706,2937
11	.005,4945	.038,4615	.148,3516	.423,0769	.670,3297
12	.004,3956	.032,9670	.129,6703	.395,6044	.637,3626
13	.003,5714	.028,5714	.114,2857	.371,4286	.607,1429
14	.002,9412	.025,0000	.101,4706	.350,0000	.579,4118
15	.002,4510	.022,0588	.090,6863	.330,8824	.553,9216
16	.002,0640	.019,6078	.081,5273	.313,7255	.530,4438
17	.001,7544	.017,5439	.073,6842	.298,2456	.508,7719
18	.001,5038	.015,7895	.066,9173	.284,2105	.488,7218
19	.001,2987	.014,2857	.061,0390	.271,4286	.470,1299
20	.001,1293	.012,9870	.055,9006	.259,7403	.452,8515

n = 4

n \ u'	2	3	4	5	6
4	.028,5714	.114,2857	.371,4286	.628,5714	.885,7143
5	.015,8730	.071,4286	.261,9048	.500,0000	.785,7143
6	.009,5238	.047,6190	.190,4762	.404,7619	.690,4762
7	.006,0606	.033,3333	.142,4242	.333,3333	.606,0606
8	.004,0404	.024,2424	.109,0909	.278,7879	.533,3333
9	.002,7972	.018,1818	.085,3147	.236,3636	.471,3287
10	.001,9580	.013,9860	.067,9321	.202,7972	.416,5814
11	.001,4652	.010,9890	.054,9451	.175,8242	.373,6264
12	.001,0989	.008,7912	.045,0549	.153,8462	.335,1648
13	.000,8403	.007,1429	.037,3950	.135,7143	.302,1008
14	.000,6536	.005,8824	.031,3725	.120,5882	.273,5294
15	.000,5160	.004,9020	.026,5738	.107,8431	.248,7100
16	.000,4128	.004,1280	.022,7038	.097,0072	.227,0382
17	.000,3342	.003,5088	.019,5489	.087,7193	.208,0200
18	.000,2734	.003,0075	.016,9515	.079,6992	.191,2509
19	.000,2259	.002,5974	.014,7939	.072,7273	.176,3975
20	.000,1882	.002,2586	.012,9870	.066,6290	.163,1846

n = 4

n \ u'	7	8
4	.971,4286	1.
5	.928,5714	.992,0635
6	.880,9524	.976,1905
7	.833,3333	.954,5455
8	.787,8788	.929,2929
9	.745,4545	.902,0979
10	.706,2937	.874,1259
11	.670,3297	.846,1538
12	.637,3626	.818,6813
13	.607,1429	.792,0168
14	.579,4118	.766,3399
15	.553,9216	.741,7441
16	.530,4438	.718,2663
17	.508,7719	.695,9064
18	.488,7218	.674,6811
19	.470,1299	.654,4325
20	.452,8515	.635,2343

TABLE I (*Continued*)

$n = 5$

n \ u'	2	3	4	5	6
5	.007,9365	.039,6825	.166,6667	.357,1429	.642,8571
6	004,3290	.023,8095	.110,3896	.261,9048	.521,6450
7	.002,5253	.015,1515	.075,7576	.196,9697	.424,2424
8	.001,5540	.010,1010	.053,6131	.151,5152	.347,3193
9	.000,9990	.006,9930	.038,9610	.118,8811	.286,7133
10	.000,6660	.004,9950	.028,9710	.094,5051	.238,7612
11	.000,4579	.003,6630	.021,9780	.076,9231	.200,5494
12	.000,3232	.002,7472	.016,9683	.063,1868	.169,8449
13	.000,2334	.002,1008	.013,3053	.052,5210	.144,9580
14	.000,1720	.001,6340	.010,5779	.044,1176	.124,6130
15	.000,1290	.001,2900	.008,5139	.037,4097	.107,8431
16	.000,0983	.001,0320	.006,9291	.031,9917	.093,9112
17	.000,0759	.000,8354	.005,6960	.027,5689	.082,2511
18	.000,0594	.000,6835	.004,7252	.023,9234	.072,4241
19	.000,0471	.000,5647	.003,9526	.020,8922	.064,0881
20	.000,0376	.000,4705	.003,3314	.018,3512	.056,9735

$n = 6$

n \ u'	2	3	4	5	6
6	.002,1645	.012,9870	.067,0996	.175,3247	.331,7749
7	.001,1655	.007,5758	.042,5408	.121,2121	.236,0373
8	.000,6660	.004,6620	.027,9720	.086,2471	.226,1072
9	.000,3996	.002,9970	.018,9810	.062,9371	.174,8252
10	.000,2498	.001,9980	.013,2368	.046,9530	.136,8631
11	.000,1616	.001,3736	.009,4538	.035,7143	.108,4357
12	.000,1077	.000,9696	.006,8951	.027,6341	.086,5886
13	.000,0737	.000,7003	.005,1231	.021,7067	.070,3597
14	.000,0516	.000,5160	.003,8700	.017,2859	.057,5335
15	.000,0369	.000,3870	.002,9670	.013,9319	.047,4716
16	.000,0268	.000,2949	.002,3052	.011,3519	.039,4971
17	.000,0198	.000,2278	.001,8128	.009,3415	.033,1154
18	.000,0149	.000,1783	.001,4414	.007,7505	.027,9652
19	.000,0113	.000,1412	.001,1575	.006,4935	.023,7719
20	.000,0087	.000,1129	.000,9382	.005,4771	.020,3318

$n = 5$

n \ u'	7	8	9	10
5	.960,3175	.992,0635	1.	
6	.911,2554	.976,1905		
7	.853,5354	.954,5455		
8	.793,3178	.929,2929		
9	.734,2657	.902,0979		
10	.678,3217	.874,1259		
11	.626,3736	.846,1538		
12	.578,7007	.818,6813		
13	.535,2474	.792,0168		
14	.495,7880	.766,3399		
15	.460,0103	.741,7441		
16	.427,5886	.718,2562		
17	.398,1924	.695,9064		
18	.371,5118	.674,6411		
19	.347,2614	.654,4325		
20	.325,1835	.635,2343		

$n = 6$

n \ u'	7	8	9	10	11	12
6	.608,2251	.824,6753	.932,9004	.987,0130	.997,8355	1.
7	.500,0000	.733,1002	.878,7879	.966,2005	.992,4242	.999,6172
8	.412,5874	.645,6876	.820,5128	.937,2629	.983,2830	.997,6690
9	.342,6573	.566,4336	.762,2378	.902,0979	.972,0280	.994,3356
10	.286,7133	.496,5035	.706,2937	.863,6364	.958,0420	.989,5105
11	.241,7552	.435,8820	.653,8462	.823,5294	.942,3077	.983,9317
12	.204,3976	.383,1610	.604,3652	.783,1285	.925,3394	.978,1131
13	.175,7703	.337,9404	.560,9244	.743,3858	.907,6630	.972,5443
14	.151,4446	.299,0196	.520,3815	.704,8904	.889,3153	.966,7279
15	.131,3209	.264,4799	.483,4881	.667,9466	.870,5720	.961,0594
16	.114,8611	.236,5137	.449,9438	.632,8824	.852,4265	.955,4204
17	.100,4765	.211,4278	.419,4577	.599,7504	.834,1308	.950,0712
18	.088,5910	.189,6342	.391,7204	.568,5149	.816,1015	.944,9508
19	.078,4867	.170,6361	.366,4596	.539,2434	.798,4130	.939,9779
20	.069,8475	.154,0242	.343,4218	.511,7752	.781,1406	.935,1525

TABLE I (*Continued*)

$s = 7$

n \ n'	2	3	4	5	6
7	.000,5828	.004,0793	.025,0583	.077,5058	.206,6247
8	.000,3108	.002,3310	.015,3816	.051,2821	.149,1841
9	.000,1746	.001,3986	.009,7902	.034,9650	.108,3916
10	.000,1028	.000,8741	.006,4274	.024,4755	.080,0082
11	.000,0628	.000,5656	.004,3363	.017,5339	.059,9548
12	.000,0397	.000,3771	.002,9967	.012,8205	.045,5664
13	.000,0258	.000,2580	.002,1156	.009,5459	.035,0677
14	.000,0172	.000,1806	.001,5222	.007,2239	.027,3478
15	.000,0117	.000,1290	.001,1141	.005,5470	.021,5546
16	.000,0082	.000,0938	.000,8280	.004,3156	.017,1645
17	.000,0058	.000,0693	.000,6241	.003,3978	.013,7993
18	.000,0042	.000,0520	.000,4764	.002,7044	.011,1920
19	.000,0030	.000,0395	.000,3679	.002,1739	.009,1517
20	.000,0023	.000,0304	.000,2872	.001,7635	.007,5403

$s = 7$

n \ n'	12	13	14
7	.995,9207	.999,4172	1.
8	.987,8788	.997,6690	.999,8446
9	.974,8252	.994,4056	.999,3007
10	.957,1164	.989,5105	.998,1489
11	.935,5204	.983,0317	.996,2293
12	.910,9312	.975,1131	.993,4508
13	.884,2105	.965,9443	.989,7833
14	.856,1146	.955,7276	.985,2825
15	.827,2704	.944,6594	.979,8762
16	.798,1783	.932,9205	.973,7515
17	.769,2254	.920,6712	.966,9464
18	.740,7031	.908,0508	.959,5423
19	.712,8246	.895,1779	.951,6206
20	.685,7403	.882,1526	.943,2587

$s = 7$

n \ n'	7	8	9	10	11
7	.383,4499	.616,5501	.791,3753	.922,4942	.974,9417
8	.296,0373	.513,5975	.703,9627	.867,1329	.948,7179
9	.230,7692	.426,5734	.622,3776	.805,9441	.916,0839
10	.181,8182	.354,5866	.548,9510	.743,3155	.879,3706
11	.144,7964	.295,6259	.484,1629	.682,1267	.840,4977
12	.116,5158	.247,4994	.427,6018	.624,0772	.800,9050
13	.094,6852	.208,2043	.378,4830	.570,0468	.761,6099
14	.077,6574	.176,0406	.335,9133	.520,1818	.723,2972
15	.064,2415	.149,6153	.298,0196	.475,1052	.686,4035
16	.053,5698	.127,8079	.264,0044	.434,0402	.651,1868
17	.045,0038	.109,7242	.234,1651	.396,9212	.617,7796
18	.038,0695	.094,6536	.208,0949	.363,4283	.586,2284
19	.032,4111	.082,0310	.185,3759	.333,2320	.556,5217
20	.027,7592	.071,4064	.165,0684	.306,0099	.528,6105

$s = 8$

n \ n'	2	3	4	5	6
8	.000,1554	.001,2432	.008,8578	.031,7016	.100,2331
9	.000,0823	.000,6993	.005,3065	.020,2797	.068,6549
10	.000,0457	.000,4114	.003,2908	.013,3690	.047,9227
11	.000,0265	.000,2514	.002,1037	.009,0498	.034,0557
12	.000,0159	.000,1588	.001,3813	.006,2713	.024,6090
13	.000,0098	.000,1032	.000,9288	.004,4376	.018,0599
14	.000,0063	.000,0688	.000,6380	.003,1992	.013,4440
15	.000,0041	.000,0469	.000,4467	.002,3454	.010,1404
16	.000,0027	.000,0326	.000,3182	.001,7458	.007,7420
17	.000,0018	.000,0231	.000,2302	.001,3175	.005,9774
18	.000,0013	.000,0166	.000,1690	.001,0069	.004,6631
19	.000,0009	.000,0122	.000,1257	.000,7784	.003,6728
20	.000,0006	.000,0090	.000,0946	.000,6081	.002,9185

TABLE I (*Continued*)

$n = 8$

n \ u'	7	8	9	10	11
8	.214,4522	.404,8174	.595,1826	.785,5478	.899,7669
9	.157,3427	.318,5932	.500,0000	.701,5631	.842,6573
10	.117,0300	.251,4055	.419,3747	.620,9379	.782,1884
11	.088,2353	.199,3729	.352,1870	.546,6778	.721,7195
12	.067,3970	.159,0855	.296,6182	.479,9952	.663,3722
13	.052,1156	.127,7950	.250,7740	.421,0526	.606,3581
14	.040,7637	.103,3712	.212,9343	.369,4530	.557,2755
15	.032,2263	.084,1930	.181,6305	.324,5390	.510,3199
16	.025,7305	.069,0368	.155,6472	.285,5639	.467,4474
17	.020,7337	.056,9771	.133,9944	.251,7856	.426,4724
18	.016,8504	.047,3188	.115,8727	.222,5120	.393,1350
19	.013,8036	.039,5324	.100,6385	.197,1217	.361,1432
20	.011,3915	.033,2151	.087,7741	.175,0684	.332,1982

$n = 8$

n \ u'	12	13	14	15	16
8	.968,2984	.991,1422	.998,7568	.999,8446	1.
9	.939,4076	.979,7203	.995,8453	.999,3007	.999,9589
10	.903,1263	.963,5952	.990,4703	.998,1489	.999,7943
11	.861,7528	.943,4389	.983,3371	.996,2293	.999,4046
12	.817,4089	.920,1000	.971,8456	.993,4508	.998,8902
13	.771,8266	.894,4272	.957,9979	.989,7833	.997,5674
14	.726,3158	.867,1827	.942,3117	.985,2425	.995,9752
15	.681,8100	.839,0093	.924,7543	.979,8762	.993,8754
16	.638,9375	.810,4276	.905,6999	.973,7515	.991,2505
17	.598,5917	.781,8459	.885,5022	.966,9464	.988,1007
18	.559,4924	.753,5760	.864,4810	.959,5423	.984,4394
19	.523,2350	.725,8498	.842,9161	.951,6206	.980,2899
20	.489,3281	.698,8345	.821,0466	.943,2587	.975,6823

$n = 9$

n \ u'	2	3	4	5	6
9	.000,0411	.000,3702	.003,0029	.012,2172	.044,4673
10	.000,0217	.000,2057	.001,7845	.007,6100	.029,4334
11	.000,0119	.000,1191	.001,0717	.004,8821	.019,8857
12	.000,0068	.000,0714	.000,6702	.003,2151	.013,6937
13	.000,0040	.000,0442	.000,4302	.002,1672	.009,4975
14	.000,0024	.000,0281	.000,2827	.001,4917	.006,8365
15	.000,0015	.000,0184	.000,1897	.001,0463	.004,9438
16	.000,0010	.000,0122	.000,1297	.000,7465	.003,6246
17	.000,0006	.000,0083	.000,0903	.000,5409	.002,6916
18	.000,0004	.000,0058	.000,0638	.000,3975	.002,0225
19	.000,0003	.000,0041	.000,0458	.000,2959	.001,5164
20	.000,0002	.000,0029	.000,0333	.000,2230	.001,1791

$n = 9$

n \ u'	7	8	9	10	11
9	.103,9675	.237,9679	.399,2184	.600,7816	.762,0321
10	.076,7174	.178,5598	.318,5932	.500,0000	.681,4068
11	.054,8940	.134,9131	.254,9417	.429,9833	.605,0250
12	.039,8905	.102,7626	.204,9097	.362,1100	.535,0083
13	.029,4118	.078,9474	.165,6347	.304,3536	.472,1352
14	.021,9814	.061,1792	.134,6749	.257,1678	.416,4057
15	.016,6363	.047,8163	.110,1763	.217,1577	.367,4115
16	.012,7388	.037,6826	.090,6888	.184,7254	.324,5310
17	.009,8606	.029,9339	.075,0988	.156,5466	.287,1228
18	.007,7099	.023,9597	.062,4530	.133,2442	.254,5038
19	.006,0849	.019,1169	.052,3969	.113,4218	.225,0867
20	.004,8444	.015,6810	.044,1269	.098,1096	.201,2567

TABLE I (*Continued*)

$n = 9$

n'	12	13	14	15	16
9	.831,0525	.955,5327	.987,7828	.996,9971	.999,6298
10	.834,1705	.923,2826	.974,2038	.992,3900	.998,6252
11	.773,0650	.885,0917	.955,1084	.985,1155	.996,5468
12	.711,0402	.843,0817	.931,1026	.975,1131	.993,0766
13	.650,4044	.799,0712	.903,0960	.962,5387	.988,0141
14	.592 7885	.754,4892	.872,0824	.947,6780	.981,2761
15	.538,9016	.710,3917	.839,0093	.930,8790	.972,8766
16	.489,1695	.667,5192	.804,7113	.912,5050	.962,9021
17	.443,6940	.626,3616	.769,5856	.892,9062	.951,4874
18	.402,3771	.587,2186	.735,0919	.872,4027	.938,7948
19	.365,0025	.550,2503	.700,7642	.851,2780	.924,9991
20	.331,7957	.515,5164	.667,2279	.829,7760	.910,2760

$n = 9$

n'	17	18
9	.999,9589	1..
10	.999,7943	.999,9892
11	.999,4046	.999,9405
12	.998,6902	.999,8129
13	.997,5674	.999,5577
14	.995,9752	.999,1251
15	.993,8754	.998,4688
16	.991,2505	.997,5501
17	.988,1007	.996,3387
18	.984,4394	.994,8131
19	.980,2899	.992,9607
20	.975,6823	.990,7760

$n = 10$

n'	2	3	4	5	6
10	.000,0108	.000,1083	.000,9851	.004,4924	.016,5217
11	.000,0057	.000,0595	.000,5689	.002,7387	.011,9246
12	.000,0031	.000,0340	.000,3402	.001,7181	.007,8420
13	.000,0017	.000,0201	.000,2069	.001,1057	.005,2593
14	.000,0010	.000,0122	.000,1315	.000,7281	.003,5016
15	.000,0006	.000,0076	.000,0847	.000,4895	.002,4939
16	.000,0004	.000,0049	.000,0557	.000,3353	.001,7586
17	.000,0002	.000,0032	.000,0373	.000,2336	.0012576
18	.000,0002	.000,0021	.000,0255	.000,1654	.000,9115
19	.000,0001	.000,0014	.000,0176	.000,1187	.000,6657
20	.000,0001	.000,0010	.000,0124	.000,0864	.000,4962

$n = 10$

n'	7	8	9	10	11
10	.051,2565	.127,6386	.242,2113	.414,0704	.585,9296
11	.034,8893	.092,0457	.184,9250	.334,9607	.500,0000
12	.024,1724	.067,0398	.142,0576	.270,6597	.424,9821
13	.017,0279	.049,3337	.109,9071	.218,9393	.360,6811
14	.012,1820	.036,6806	.085,6777	.177,5474	.306,1650
15	.008,8413	.027,5493	.067,3038	.144,4744	.260,2302
16	.006,5028	.020,8935	.053,2728	.118,0313	.221,6449
17	.004,8423	.015,9941	.042,4797	.096,8449	.189,2657
18	.003,6475	.012,3527	.034,1158	.079,8184	.162,0831
19	.002,7769	.009,6211	.027,5869	.066,0851	.139,7318
20	.002,1353	.007,5516	.022,4538	.054,9634	.119,9827

$n = 10$

n'	12	13	14	15	16
10	.757,4887	.872,3614	.948,7432	.981,4783	.995,5076
11	.680,0429	.815,0750	.915,0283	.965,1107	.989,6064
12	.605,0250	.755,0607	.875,0593	.943,6771	.980,4205
13	.535,1326	.695,0464	.830,7309	.917,9567	.967,7999
14	.471,5305	.636,8959	.783,8875	.888,8814	.951,8778
15	.414,5713	.581,7741	.736,1152	.857,3832	.932,9759
16	.364,1136	.530,3271	.688,6257	.824,3101	.911,4359
17	.319,7420	.482,8375	.642,3057	.790,3690	.888,0244
18	.280,9096	.439,3464	.597,7810	.756,2166	.862,9182
19	.247,0268	.399,7345	.555,4405	.722,2661	.836,6609
20	.217,5115	.363,9048	.515,5184	.688,9010	.809,5510

TABLE I (*Continued*)

$n = 10$

n \ a'	17	18	19	20
10	.999,0149	.999,8917	.999,9892	1.
11	.997,2613	.999,5577	.999,9405	.999,9972
12	.994,1993	.998,7922	.999,9129	.999,9830
13	.989,5264	.997,3944	.999,5577	.999,9423
14	.983,3760	.995,1878	.999,1251	.999,8542
15	.975,5014	.992,0380	.998,4688	.999,6938
16	.966,0520	.987,8584	.997,5501	.999,4346
17	.955,1487	.982,6087	.996,3387	.999,0508
18	.942,9443	.976,2885	.994,8131	.998,5180
19	.929,6066	.968,9298	.992,9607	.997,8154
20	.915,3073	.960,5885	.990,7760	.996,9253

$n = 11$

n \ a'	2	3	4	5	6
11	.000,0028	.000,0312	.000,3147	.001,5905	.007,3317
12	.000,0015	.000,0170	.000,1797	.000,9526	.004,6136
13	.000,0008	.000,0096	.000,1058	.000,5865	.002,9662
14	.000,0004	.000,0056	.000,0639	.000,3702	.001,9451
15	.000,0003	.000,0034	.000,0396	.000,2389	.001,2990
16	.000,0002	.000,0021	.000,0251	.000,1574	.000,8822
17	.000,0001	.000,0013	.000,0162	.000,1056	.000,6085
18	.000,0001	.000,0008	.000,0107	.000,0721	.000,4259
19	.000,0000	.000,0005	.000,0071	.000,0500	.000,3020
20	.000,0000	.000,0004	.000,0049	.000,0351	.000,2169

$n = 11$

n \ a'	7	8	9	10	11
11	.022,6414	.063,4675	.134,9131	.259,9426	.409,9786
12	.014,9866	.044,2748	.099,1903	.201,6992	.334,9607
13	.010,1052	.031,2578	.073,5631	.156,8515	.273,4554
14	.006,9323	.022,3314	.055,0545	.122,4256	.223,4623
15	.004,8324	.016,1395	.041,5803	.095,9954	.183,0596
16	.003,4190	.011,7946	.031,6866	.075,6585	.150,4106
17	.002,4526	.008,7113	.024,3580	.059,9542	.124,0275
18	.001,7820	.006,4992	.018,8817	.047,7741	.102,6697
19	.001,3103	.004,8954	.014,7542	.038,2809	.085,3343
20	.000,9742	.003,7208	.011,6173	.030,8434	.071,2182

$n = 11$

n \ a'	12	13	14	15	16
11	.590,0214	.740,0572	.865,0869	.936,5325	.977,3586
12	.507,1756	.665,0393	.808,5517	.900,8097	.959,3862
13	.433,4892	.593,2831	.748,7549	.859,6062	.935,9556
14	.369,0039	.526,6523	.688,3430	.815,3856	.907,7803
15	.313,6559	.466,0183	.629,2637	.769,1883	.875,7976
16	.266,4962	.411,6032	.572,8332	.722,5468	.841,0015
17	.226,5446	.363,2342	.519,8576	.676,4811	.804,3369
18	.192,8141	.320,5187	.470,7594	.631,7315	.766,6415
19	.164,3540	.282,9585	.425,6872	.588,8056	.728,6214
20	.140,4323	.250,0211	.384,6040	.548,0260	.690,8486

$n = 11$

n \ a'	17	18	19	20	21
11	.992,6683	.998,4095	.999,6853	.999,9688	.999,9972
12	.985,0134	.995,9965	.999,0474	.999,8610	.999,9830
13	.974,0303	.9918779	.997,8270	.999,5898	.999,9423
14	.959,7523	.985,7383	.995,8440	.999,0521	.999,8542
15	.942,4283	.977,4095	.992,9567	.998,1390	.999,6938
16	.922,4391	.966,8596	.989,0699	.996,7475	.999,4346
17	.900,2288	.954,1680	.984,1342	.994,7889	.999,0508
18	.876,2559	.939,4949	.978,1410	.992,1941	.998,5180
19	.850,9602	.923,0528	.971,1144	.988,9151	.997,8154
20	.824,7448	.905,0825	.963,1042	.984,9243	.996,9253

$n = 11$

n \ a'	22
11	1.
12	.999,9993
13	.999,9952
14	.999,9825
15	.999,9529
16	.999,8953
17	.999,7966
18	.999,6423
19	.999,4174
20	.999,1074

TABLE I (*Continued*)

$n = 12$

n'	2	3	4	5	6
12	.000,0007	.000,0089	.000,0984	.000,5458	.002,7831
13	.000,0004	.000,0048	.000,0556	.000,3221	.001,7182
14	.000,0002	.000,0027	.000,0323	.000,1952	.001,0836
15	.000,0001	.000,0016	.000,0193	.000,1211	.000,6970
16	.000,0001	.000,0009	.000,0118	.000,0769	.000,4565
17	.000,0000	.000,0006	.000,0073	.000,0497	.000,3041
18	.000,0000	.000,0003	.000,0047	.000,0328	.000,2057
19	.000,0000	.000,0002	.000,0030	.000,0220	.000,1412
20	000,0000	.000,0001	.000,0020	.000 0150	.000 0983

$n = 12$

n'	7	8	9	10	11
12	.009,4950	.029,6307	.069,9020	.150,4447	.263,2045
13	.006,1391	.020,0938	.049,7664	.112,5897	.206,8246
14	.004,0450	.013,8175	.035,8056	.084,6662	.162,8483
15	.002,7123	.009,6222	.026,0331	.064,0373	.128,6445
16	.001,8486	.006,7842	.019,1233	.048,7369	.102,0415
17	.001,2791	.004,8401	.014,1876	.037,3340	.081,3120
18	.000,8976	.003,4920	.010,6267	.028,7876	.065,1096
19	.000,6381	.002,5463	.008,0322	.022,3434	.052,3969
20	.000,4593	.001,8755	.006,1241	.017,4538	.042,3791

$n = 12$

n'	12	13	14	15	16
12	.421,0682	.578,9318	.736,7955	.849,5553	.930,0980
13	.347,5488	.500,0000	.664,1782	.793,1754	.893,6927
14	.285,9820	.429,6379	.593,8161	.734,5403	.851,8105
15	.235,0563	.368,0711	.527,6888	.675,9052	.806,2054
16	.193,2517	.314,8652	.466,8820	.618,8989	.758,5062
17	.159,0836	.269,2601	.411,8414	.564,8072	.710,0983
18	.131,2154	.230,3743	.362,5861	.513,6853	.662,0863
19	.108,4967	.197,3213	.318,8709	.466,4668	.615,3029
20	.089,9637	.169,2714	.280,3023	.423,0562	.570,3420

$n = 12$

n'	17	18	19	20	21
12	.970,3693	.990,5050	.997,2169	.999,4542	.999,9016
13	.950,2336	.981,6453	.993,8609	.998,9245	.999,6779
14	.925,1043	.969,0806	.988,6257	.996,7694	.999,2125
15	.895,7868	.952,7931	.981,2963	.993,9643	.998,3982
16	.863,2217	.933,0154	.971,7952	.989,8925	.997,1314
17	.828,3099	.910,1487	.960,1613	.984,4098	.995,3216
18	.791,9371	.884,6877	.946,5215	.977,4383	.992,8968
19	.754,8369	.857,1617	.931,0630	.968,9611	.989,8051
20	.717,6273	.828,0922	.914,0089	.959,0129	.986,0153

$n = 12$

n'	22	23	24
12	.999,9911	.999,9993	1.
13	.999,9571	.999,9952	.999,9998
14	.999,8640	.999,9825	.999,9987
15	.999,6650	.999,9529	.999,9948
16	.999,3030	.999,8953	.999,9850
17	.998,7164	.999,7966	.999,9649
18	.997,8435	.999,6423	.999,9285
19	.996,6268	.999,4174	.999,8685
20	.995,0161	.999,1074	.999,7768

$n = 13$

n'	2	3	4	5	6
13	.000,0002	.000,0025	.000,0302	.000,1825	.001,0201
14	.000,0001	.000,0013	.000,0169	.000,1063	.000,6196
15	.000,0001	.000,0007	.000,0097	.000,0636	.000,3844
16	.000,0000	.000,0004	.000,0057	.000,0389	.000,2431
17	.000,0000	.000,0003	.000,0035	.000,0243	.000,1566
18	.000,0000	.000,0002	.000,0021	.000,0155	.000,1025
19	.000,0000	.000,0001	.000,0013	.000,0100	.000,0682
20	.000,0000	.000,0001	.000,0009	.000,0066	.000,0460

TABLE I (*Continued*)

$n = 13$

$n^{n'}$	7	8	9	10	11
13	.003,8123	.013,1194	.034,0605	.081,1780	.156,5660
14	.002,4162	.008,6899	.023,5900	.058,8796	.118,8720
15	.001,5607	.005,8382	.016,5321	.042,9993	.090,6403
16	.001,0260	.003,9760	.011,7198	.031,6825	.069,4665
17	.000,6856	.002,7431	.008,4011	.025,4462	.053,5364
18	.000,4652	.001,9158	.006,0854	.017,5103	.041,5083
19	.000,3201	.001,3537	.004,4545	.013,1753	.032,3612
20	.000,2232	.000,9671	.003,2917	.009,9865	.025,3845

$n = 13$

$n^{n'}$	12	13	14	15	16
13	.277,1867	.417,9109	.582,0891	.722,8133	.843,4340
14	.220,5061	.347,5488	.505,6463	.652,4512	.787,9634
15	.175,3354	.288,8822	.436,4787	.584,6971	.729,8867
16	.139,5990	.238,8567	.375,1477	.521,1736	.671,3720
17	.111,3096	.197,9694	.321,5399	.462,7634	.614,0742
18	.089,0234	.164,2684	.275,1557	.409,8046	.559,1630
19	.071,4307	.136,5466	.235,3056	.362,2814	.507,3967
20	.057,5196	.113,7560	.201,2349	.319,9562	.459,2082

$n = 13$

$n^{n'}$	17	18	19	20	21
13	.918,8220	.965,9395	.986,8806	.996,1877	.998,9799
14	.881,1280	.948,6493	.976,4100	.992,0943	.997,5838
15	.838,7805	.918,1821	.962,2942	.985,8206	.995,2312
16	.793,4061	.887,2820	.944,6493	.977,0996	.991,7022
17	.746,4711	.852,8615	.923,7885	.965,8193	.986,8347
18	.699,1865	.815,8728	.900,1462	.952,0067	.980,5300
19	.652,5120	.777,2204	.874,2139	.935,8003	.972,7510
20	.607,1635	.737,7122	.846,5028	.917,4183	.963,5133

$n = 13$

$n^{n'}$	22	23	24	25	26
13	.999,8175	.999,9698	.999,9975	.999,9998	1.
14	.999,4659	.999,8937	.999,9870	.999,9987	.999,9999
15	.998,7601	.999,7226	.999,9559	.999,9948	.999,9996
16	.997,5432	.999,4017	.999,8845	.999,9850	.999,9985
17	.995,6611	.998,8707	.999,7461	.999,9649	.999,9953
18	.992,9766	.998,0683	.999,5084	.999,9285	.999,9885
19	.989,3787	.996,9368	.999,1355	.999,8685	.999,9753
20	.984,7879	.995,4252	.998,5900	.999,7768	.999,9527

$n = 14$

$n^{n'}$	2	3	4	5	6
14	.000,0000	.000,0007	.000,0095	.000,0597	.000,3630
15	.000,0000	.000,0004	.000,0051	.000,0344	.000,2174
16	.000,0000	.000,0002	.000,0029	.000,0203	.000,1330
17	.000,0000	.000,0001	.000,0017	.000,0123	.000,0809
18	.000,0000	.000,0001	.000,0010	.000,0076	.000,0526
19	.000,0000	.000,0000	.000,0006	.000,0048	.000,0339
20	.000,0000	.000,0000	.000,0004	.000,0030	.000,0222

$n = 14$

$n^{n'}$	7	8	9	10	11
14	.001,4751	.005,5531	.015,7478	.041,2348	.087,1113
15	.000,9191	.003,6036	.010,6505	.029,1065	.064,1731
16	.000,5835	.002,3732	.007,2948	.020,7174	.047,5626
17	.000,3770	.001,5849	.005,0577	.014,8721	.035,4823
18	.000,2476	.001,0726	.003,5478	.010,7670	.026,6453
19	.000,1651	.000,7351	.002,5165	.007,8606	.020,1530
20	.000,1116	.000,5098	.001,8039	.005,7858	.015,3423

TABLE I (*Continued*)

$n = 14$

n \ u'	12	13	14	15	16
14	.169,6891	.279,7927	.426,5976	.573,4024	.720,2073
15	.130,6149	.224,7409	.357,6246	.500,0000	.651,8671
16	.100,7161	.180,4463	.298,5652	.433,5551	.585,4252
17	.077,8803	.145,0107	.248,6504	.374,4937	.522,5555
18	.060,4353	.116,7453	.206,8412	.322,6788	.464,2682
19	.047,0863	.094,2213	.172,0314	.277,5305	.411,2197
20	.036,8445	.076,2651	.143,1807	.238,7258	.363,9605

$n = 14$

n \ u'	17	18	19	20	21	22
14	.830,3109	.912,8887	.958,7652	.984,2522	.994,4469	.998,5248
15	.775,2591	.874,9219	.935,8269	.972,7391	.989,3495	.996,7320
16	.721,3090	.832,2093	.908,1428	.957,3590	.981,9671	.993,7790
17	.661,3588	.786,2817	.876,5039	.938,1942	.972,1239	.989,3972
18	.605,8375	.738,5681	.841,8030	.915,5423	.959,7858	.983,3824
19	552,7452	.690,3024	.804,9334	.889,8453	.945,0380	.975,6062
20	.502,7244	.642,4884	.766,7231	.861,6246	.928,0556	.966,0162

$n = 14$

n \ u'	23	24	25	26	27	28
14	.999,6370	.999,9403	.999,9909	.999,9993	.999,9999	1.
15	.999,0809	.999,8131	.999,9656	.999,9961	.999,9998	.999,99999
16	.998,0742	.999,5385	.999,9046	.999,9859	.999,9985	.999,9999
17	.996,4535	.999,0331	.999,7826	.999,9610	.999,9953	.999,9995
18	.994,1081	.998,2033	.999,5684	.999,9097	.999,9885	.999,9986
19	.990,8904	.996,9535	.999,2272	.999,8166	.999,9753	.999,9963
20	.986,7220	.995,1925	.998,7219	.999,6631	.999,9527	.999,9916

$n = 15$

n \ u'	2	3	4	5	6
15	.000,0000	.000,0002	.000,0027	.000,0131	.000,1259
16	.000,0000	.000,0001	.000,0015	.000,0109	.000,0745
17	.000,0000	.000,0001	.000,0008	.000,0064	.000,0450
18	.000,0000	.000,0000	.000,0005	.000,0038	.000,0277
19	.000,0000	.000,0000	.000,0003	.000,0023	.000,0173
20	.000,0000	.000,0000	.000,0002	.000,0014	.000,0110

$n = 15$

n \ u'	7	8	9	10	11
15	.000,5530	.002,2613	.006,9592	.019,3785	.045,7170
16	.000,3395	.001,4416	.004,6103	.013,7030	.032,7977
17	.000,2123	.000,9329	.003,0948	.009,5355	.023,7050
18	.000,1351	.000,6124	.002,1040	.006,6980	.017,2643
19	.000,0873	.000,4074	.001,4477	.004,7484	.012,6703
20	.000,0573	.000,2745	.001,0076	.003,3967	.009,3695

$n = 15$

n \ u'	12	13	14	15	16
15	.097,3940	.174,9095	.291,1827	.424,0664	.575,9336
16	.072,8057	.136,1517	.236,1717	.357,6246	.504,5928
17	.054,6203	.106,1457	.191,1627	.300,4703	.439,2736
18	.041,1534	.082,9583	.154,6265	.251,8392	.380,5976
19	.031,1546	.065,2425	.125,1165	.210,9365	.328,6325
20	.023,7043	.051,1792	.101,3509	.176,6085	.283,0963

TABLE I (*Continued*)

$n = 15$

n \ u'	17	18	19	20	21
15	.708,8173	.825,0905	.902,6060	.954,2830	.980,1215
16	.642,3754	.770,9725	.863,8483	.930,5283	.967,2023
17	.578,0765	.714,7113	.820,9825	.901,9511	.950,5323
18	.517,3504	.658,1253	.775,4377	.869,2877	.930,2501
19	.461,0404	.602,6434	.728,5128	.833,4039	.906,8276
20	.409,5455	.549,3124	.681,3118	.795,1935	.880,6049

$n = 15$

n \ u'	22	23	24	25	26
15	.993,0408	.997,7387	.999,4470	.999,8741	.999,9809
16	.987,2063	.995,3897	.998,6962	.999,6605	.999,9361
17	.978,8713	.991,7526	.997,3736	.999,2472	.999,8327
18	.967,8301	.986,6001	.995,2870	.998,5446	.999,6305
19	.954,0286	.979,7746	.992,2575	.997,4567	.999,2792
20	.937,5457	.971,1926	.988,1337	.995,8984	.998,7215

$n = 15$

n \ u'	27	28	29	30
15	.999,9973	.999,9998	.999,9999	1.
16	.999,9891	.999,9988	.999,9999	1.
17	.999,9679	.999,9956	.999,9995	1.
18	.999,9228	.999,9871	.999,9986	.999,9999
19	.999,8393	.999,9686	.999,9963	.999,9996
20	.999,6993	.999,9332	.999,9916	.999,9988

$n = 16$

n \ u'	2	3	4	5	6
16	.000,0000	.000,0001	.000,0008	.000,0060	.000,0427
17	.000,0000	.000,0000	.000,0004	.000,0034	.000,0250
18	.000,0000	.000 0000	.000,0002	.000,0020	.000,0149
19	.000,0000	.000,0000	.000,0001	.000,0012	.000,0091
20	.000,0000	.000,0000	.000,0001	.000,0007	.000,0056

$n = 16$

n \ u'	7	8	9	10	11
16	.000,2017	.000,8905	.002,9571	.009,1566	.022,7957
17	.000 1222	000,5530	001,9235	.006,1821	.015,9762
18	.000,0754	.000,3562	.001,2687	.004,2167	.011,2921
19	.000,0473	.000,2302	.000,8475	.002,9051	.008,0492
20	.000,0302	000,1500	.000,573?	.002,0211	.005,7858

$n = 16$

n \ u'	12	13	14	15	16
16	.052,8017	.102,8117	.186,1617	.293,3260	.431,1?57
17	.038,4?00	.077,8067	.145,6073	.239,7439	.365,9287
18	.028,1549	.059,0702	.115,2798	.195,5792	.309,1465
19	.020,7242	.045,0178	.090,7885	.159,4415	.260,3257
20	.015,3423	.034,4554	.071,5196	.130,0205	.218,7595

TABLE I (*Continued*)

$m = 16$

n \ u'	17	18	19	20	21
16	.568,8913	.706,6740	.813 8383	.897,1883	.947,1983
17	.500,0000	.641 9579	.750,2561	858,3398	.922 1933
18	436 9076	.578,8556	.705,0503	.815,4620	892,7501
19	.380,1245	.518 8376	649 8414	.769,7200	.859 6266
20	.329,6833	.462 7919	.595,9004	.722,4357	.823,6840

$m = 16$

n \ u'	22	23	24	25	26
16	.977,2043	.990,8433	.997 0429	.999,1095	.999 7983
17	.963 4136	.984,0238	.994 2437	.998,0762	.999,4956
18	.945,7478	.974,6555	.989,9854	.996,3729	.998,9279
19	.924,3594	.962,6106	.9844,0099	.993,8179	.997,9769
20	.899,5852	947,8986	.976,1338	.990,2514	.996,5258

$m = 16$

n \ u'	27	28	29	30	31	32
16	.999,9573	.999,9940	.999 9992	.999,9999	1.	1.
17	.999,8778	.999,9786	.999,9926	.999,9997	1.	1
18	.999,7140	.999,9408	.999,9894	.999,9986	.999,9999	1.
19	.999,4192	.999 8624	.999,9732	.999,9958	.999,9996	1.
20	.998,9391	.999,7188	.999,9415	.999,9893	.999,9985	.999,9999

$m = 17$

n \ u'	2	3	4	5	6
17	.000,0000	.000,0000	.000,0002	.000,0019	.000,0142
18	.000,0000	.000,0000	000,0001	.000,0011	.000,0083
19	.000,0000	.000,0000	.000,0001	.000,0006	.000,0049
20	.000 0000	.000,0000	.000,0000	.000,0004	.000,0029

$m = 17$

n \ u'	7	8	9	10	11
17	.000,0718	.000,3406	.001,2141	.004,0520	010,8662
18	.000,0430	.000,2109	.000,7773	002,6855	.007,4596
19	.000,0262	.000,1325	.000,5046	.001 8001	.005,1685
20	.000,0163	.000 0845	000,3318	.001,2189	.003,6139

$m = 17$

n \ u'	12	13	14	15	16
17	.027,2151	.057,1966	.112,1570	.190,6720	.302 8353
18	.019,3731	.042 2073	.085,8302	.151 4145	2[illegible] 1781
19	.013,8746	031,8656	.065,8589	.120,2125	.204,9038
20	.010,0005	.023,3061	.050,6268	.095,5108	157,[illegible]44

$m = 17$

n \ u'	17	18	19	20	21
17	.429,0211	.570,9789	697 1637	.809,3280	.887,8430
18	.365,9287	.503,8306	.634,0713	.756,5508	.[illegible]49,[illegible]85[illegible]
19	.310,7679	.441,7747	.572,7816	.702 1711	.8[illegible],5826
20	.263,1290	.385,4450	.514,5763	.[illegible]47,4428	.760,3963

TABLE I (*Continued*)

$n = 17$

n \ u	22	23	24	25	26
17	.942,8034	.972,7819	.989,[illegible]18	.995,9471	.998,7859
18	.917,2300	.957,7927	[illegible],8197	.992,5404	.997,5044
19	.857,1980	.939,6718	.971,4081	.987,5764	.995,4361
20	.853,4189	[illegible],6400	.968,3533	.980,8396	.992,3711

$n = 17$

n \ u	27	28	29	30	31
17	.999,6594	.999,9282	.999,9858	.999,9981	.999,9998
18	.999,2227	.999,8101	.999,9570	.999,9929	.999,9989
19	.998,4590	.999,5751	.999,8940	.999,9795	.999,9965
20	.997,2495	.999,1604	.999,7745	.999,9499	.999,9909

$n = 17$

n \ u	32	33	34
17	1.	1.	1.
18	.999,9999	1.	1.
[illegible]	[illegible]	1.	1.
[illegible]	[illegible]	[illegible]	[illegible]

$n = 18$

n \ u	2	3	4	5	6
18	.000,0000	.000,0000	.000,0001	.000,0006	.000,0047
19	.000,0000	.000,0000	.000,0000	.000 0003	.000,0027
20	.000,0000	.000,0000	.000,0000	.000,0002	.000,0015

$n = 18$

n \ u	7	8	9	10	11
18	.000,0250	.000,1269	.000,4836	.001,7319	.004,9776
19	.000,0148	.000,0776	.000,3053	.001,1295	.003,3548
20	.000,0090	.000,0482	.000,1964	.000,7448	.002,2833

$n = 18$

n \ u	12	13	14	15	16
18	.013 4163	.030,2938	.064,0488	.117,0923	.200,4463
19	.009,3549	.021,8551	.047,3555	040,5705	.160,5125
20	006,5691	.015,8550	.035,8463	.070,1416	.128,5099

$n = 18$

n \ u	17	18	19	20	21
18	.304,6389	.434,8796	.565,1204	.605,3611	.799,5537
19	.252,5426	.372,9273	.500,0000	.633,7607	.747,4574
20	.208,7684	.318,2070	.439,8077	.573,5684	.693,9531

$n = 18$

n \ u	22	23	24	25	26
18	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

$n = 18$

n \ u	27	28	29	30	31
18	.998,2681	.999,5164	.999,8731	.999,9750	.999,9954
19	.996,6452	.998,9530	.999,6847	.999,9302	.999,9852
20	.994,1452	.997,9914	.999,3561	.999,8360	.999,9616

$n = 18$

n \ u	32	33	34	35	36
18	.999,3994	.999,9999	1.	1.	1.
19	.999,9977	.999,9997	1.	1.	1.
20	999,9930	.999,9989	.999,9998	.999,99999	1.

TABLE I (*Concluded*)

n = 19

n \ u'	2	3	4	5	6
19	.000,0000	.000,0000	.000,0000	.000,0002	.000,0015
20	.000,0000	.000,0000	.000,0000	.000,0001	.000,0009

n = 19

n \ u'	7	8	9	10	11
19	.000,0086	.000,0462	.000,1875	.000,7174	.002,2009
20	.000,0050	.000,0280	.000,1169	.000,4611	.001,4591

n = 19

n \ u'	12	13	14	15	16
19	.006,3448	.016,3650	.034,8553	.068,2844	.125,5915
20	.004,3501	.010,8549	.025,4705	.051,5699	.098,1013

n = 19

n \ u'	17	18	19	20	21
19	.204,3888	.312,7350	.433,1196	.566,8804	.687,2850
20	.164,9901	.260,9611	.372,9273	.503,2583	.627,0727

n = 19

n \ u'	22	23	24	25	26
19	.795,6112	.874,4085	.931,7156	.965,1447	.984,5450
20	.744,7706	.835,0059	.901,8070	.948,4301	.975,5774

n = 19

n \ u'	27	28	29	30	31
19	.993,6152	.997,7991	.999,2826	.999,8125	.999,9538
20	.989,1151	.995,8998	.998,5409	.999,5734	.999,8331

n = 19

n \ u'	32	33	34	35	36
19	.999,9914	.999,9985	.999,9998	.999,9999	1.
20	.999,9749	.999,9950	.999,9993	.999,9999	1.

n = 20

n \ u'	2	3	4	5	6
20	.000,0000	.000,0000	.000,0000	.000,0001	.000,0005

n = 20

n \ u'	7	8	9	10	11
20	.000,0029	.000,0165	.000,0710	.000,2890	.000,9429

n = 20

n \ u'	12	13	14	15	16
20	.002,9046	.007,4821	.018,1627	.037,9982	.074,8356

n = 20

n \ u'	17	18	19	20	21
20	.130,0916	.212,9756	.314,2744	.438,0928	.561,9072

n = 20

n \ u'	22	23	24	25	26
20	.685,7216	.787,0244	.869,9084	.925,1644	.962,0018

n = 20

n \ u'	27	28	29	30	31
20	.981,3373	.992,5179	.997,0954	.999,057	.999,7110

n = 20

n \ u'	32	33	34	35	36
20	.999,9290	.999,9835	.999,9971	.999,9995	.999,9999

n = 20

n \ u'	37	38	39	40
20	1.	1.	1.	1.

TABLE II

Significance Levels of u

Where $\varepsilon < .50$, u_ε is the *largest* integer, u', for which

$$P\{u \leq u'\} \leq \varepsilon .$$

When $\varepsilon > .50$, u is the *smallest* integer, u', for which

$$P\{u \leq u'\} \geq \varepsilon .$$

$n = 2$

n	$u_{.005}$	$u_{.01}$	$u_{.025}$	$u_{.05}$	$u_{.95}$	$u_{.975}$	$u_{.99}$	$u_{.995}$
2					4	4	4	4
3					5	5	5	5
4					5	5	5	5
5					5	5	5	5
6					5	5	5	5
7					5	5	5	5
8				2	5	5	5	5
9				2	5	5	5	5
10				2	5	5	5	5
11				2	5	5	5	5
12			2	2	5	5	5	5
13			2	2	5	5	5	5
14			2	2	5	5	5	5
15			2	2	5	5	5	5
16			2	2	5	5	5	5
17			2	2	5	5	5	5
18			2	2	5	5	5	5
19		2	2	2	5	5	5	5
20		2	2	2	5	5	5	5

$n = 3$

n	$u_{.005}$	$u_{.01}$	$u_{.025}$	$u_{.05}$	$u_{.95}$	$u_{.975}$	$u_{.99}$	$u_{.995}$
3					6	6	6	6
4					7	7	7	7
5				2	7	7	7	7
6			2	2	7	7	7	7
7			2	2	7	7	7	7
8			2	2	7	7	7	7
9		2	2	2	7	7	7	7
10		2	2	3	7	7	7	7
11		2	2	3	7	7	7	7
12	2	2	2	3	7	7	7	7
13	2	2	2	3	7	7	7	7
14	2	2	2	3	7	7	7	7
15	2	2	3	3	7	7	7	7
16	2	2	3	3	7	7	7	7
17	2	2	3	3	7	7	7	7
18	2	2	3	3	7	7	7	7
19	2	2	3	3	7	7	7	7
20	2	2	3	3	7	7	7	7

$n = 4$

n	$u_{.005}$	$u_{.01}$	$u_{.025}$	$u_{.05}$	$u_{.95}$	$u_{.975}$	$u_{.99}$	$u_{.995}$
4				2	7	8	8	8
5			2	2	8	8	8	9
6		2	2	3	8	8	9	9
7		2	2	3	8	9	9	9
8	2	2	3	3	9	9	9	9
9	2	2	3	3	9	9	9	9
10	2	2	3	3	9	9	9	9
11	2	2	3	3	9	9	9	9
12	2	3	3	4	9	9	9	9
13	2	3	3	4	9	9	9	9
14	2	3	3	4	9	9	9	9
15	3	3	4	4	9	9	9	9
16	3	3	4	4	9	9	9	9
17	3	3	4	4	9	9	9	9
18	3	3	4	4	9	9	9	9
19	3	3	4	4	9	9	9	9
20	3	3	4	4	9	9	9	9

$n = 5$

n	$u_{.005}$	$u_{.01}$	$u_{.025}$	$u_{.05}$	$u_{.95}$	$u_{.975}$	$u_{.99}$	$u_{.995}$
5		2	2	3	8	9	9	10
6	2	2	3	3	9	9	10	10
7	2	2		3	9	10	10	11
8	2	2		3	10	10	11	11
9	2	3	3	4	10	11	11	11
10	3	3	3	4	10	11	11	11
11	3	3	4	4	11	11	11	11
12	3	3	4	4	11	11	11	11
13	3	3	4	4	11	11	11	11
14	3	3	4	5	11	11	11	11
15	3	4	4	5	11	11	11	11
16	3	4	4	5	11	11	11	11
17	3	4	4	5	11	11	11	11
18	4	4	5	5	11	11	11	11
19	4	4	5	5	11	11	11	11
20	4	4	5	5	11	11	11	11

TABLE II (*Continued*)

$m = 6$

n	$u_{.005}$	$u_{.01}$	$u_{.025}$	$u_{.05}$	$u_{.95}$	$u_{.975}$	$u_{.99}$	$u_{.995}$
6	2	2	3	3	10	10	11	11
7	2	3	3	4	10	11	11	12
8	3	3	3	4	11	11	12	12
9	3	3	4	4	11	12	12	13
10	3	3	4	5	11	12	13	13
11	3	4	4	5	12	12	13	13
12	3	4	4	5	12	12	13	13
13	3	4	5	5	12	13	13	13
14	4	4	5	5	12	13	13	13
15	4	4	5	6	13	13	13	13
16	4	5	5	6	13	13	13	13
17	4	5	5	6	13	13	13	13
18	4	5	5	6	13	13	13	13
19	4	5	6	6	13	13	13	13
20	4	5	6	6	13	13	13	13

$m = 7$

n	$u_{.005}$	$u_{.01}$	$u_{.025}$	$u_{.05}$	$u_{.95}$	$u_{.975}$	$u_{.99}$	$u_{.995}$
7	3	3	3	4	11	12	12	12
8	3	3	4	4	12	12	13	13
9	3	4	4	5	12	13	13	14
10	3	4	5	5	12	13	14	14
11	4	4	5	5	13	13	14	14
12	4	4	5	6	13	13	14	15
13	4	5	5	6	13	14	15	15
14	4	5	6	6	13	14	15	15
15	4	5	6	6	14	14	15	15
16	5	5	6	6	14	15	15	15
17	5	5	6	7	14	15	15	15
18	5	5	6	7	14	15	15	15
19	5	6	6	7	14	15	15	15
20	5	6	6	7	14	15	15	15

$m = 8$

n	$u_{.005}$	$u_{.01}$	$u_{.025}$	$u_{.05}$	$u_{.95}$	$u_{.975}$	$u_{.99}$	$u_{.995}$
8	3	4	4	5	12	13	13	14
9	3	4	5	5	13	13	14	14
10	4	4	5	6	13	14	14	15
11	4	5	5	6	14	14	15	15
12	4	5	6	6	14	15	15	16
13	5	5	6	6	14	15	16	16
14	5	5	6	7	15	15	16	16
15	5	5	6	7	15	15	16	17
16	5	6	6	7	15	16	16	17
17	5	6	7	7	15	16	17	17
18	6	6	7	8	15	16	17	17
19	6	6	7	8	15	16	17	17
20	6	6	7	8	16	16	17	17

$m = 9$

n	$u_{.005}$	$u_{.01}$	$u_{.025}$	$u_{.05}$	$u_{.95}$	$u_{.975}$	$u_{.99}$	$u_{.995}$
9	4	4	5	6	13	14	15	15
10	4	5	5	6	14	15	15	16
11	5	5	6	6	14	15	16	16
12	5	5	6	7	15	15	16	17
13	5	6	6	7	15	16	17	17
14	5	6	7	7	16	16	17	17
15	6	6	7	8	16	17	17	18
16	6	6	7	8	16	17	17	18
17	6	7	7	8	16	17	18	18
18	6	7	8	8	17	17	18	18
19	6	7	8	8	17	17	18	19
20	7	7	8	9	17	17	18	19

TABLE II (*Continued*)

$n = 10$

n	$u_{.005}$	$u_{.01}$	$u_{.025}$	$u_{.05}$	$u_{.95}$	$u_{.975}$	$u_{.99}$	$u_{.995}$
10	5	5	6	6	15	15	16	16
11	5	5	6	7	15	16	17	17
12	5	6	7	7	16	16	17	18
13	5	6	7	8	16	17	18	18
14	6	6	7	8	16	17	18	18
15	6	7	7	8	17	17	18	19
16	6	7	8	8	17	18	19	19
17	7	7	8	9	17	18	19	19
18	7	7	8	9	18	18	19	20
19	7	8	8	9	18	19	19	20
20	7	8	9	9	18	19	19	20

$n = 11$

n	$u_{.005}$	$u_{.01}$	$u_{.025}$	$u_{.05}$	$u_{.95}$	$u_{.975}$	$u_{.99}$	$u_{.995}$
11	5	6	7	7	16	16	17	18
12	6	6	7	8	16	17	18	18
13	6	6	7	8	17	18	18	19
14	6	7	8	8	17	18	19	19
15	7	7	8	9	18	18	19	20
16	7	7	8	9	18	19	20	20
17	7	8	9	9	18	19	20	21
18	7	8	9	10	19	19	20	21
19	8	8	9	10	19	20	21	21
20	8	8	9	10	19	20	21	21

$n = 12$

n	$u_{.005}$	$u_{.01}$	$u_{.025}$	$u_{.05}$	$u_{.95}$	$u_{.975}$	$u_{.99}$	$u_{.995}$
12	6	7	7	8	17	18	18	19
13	6	7	8	9	17	18	19	20
14	7	7	8	9	18	19	20	20
15	7	8	8	9	18	19	20	21
16	7	8	9	10	19	20	21	21
17	8	8	9	10	19	20	21	21
18	8	8	9	10	20	20	21	22
19	8	9	10	10	20	21	22	22
20	8	9	10	11	20	21	22	22

$n = 13$

n	$u_{.005}$	$u_{.01}$	$u_{.025}$	$u_{.05}$	$u_{.95}$	$u_{.975}$	$u_{.99}$	$u_{.995}$
13	7	7	8	9	18	19	20	20
14	7	8	9	9	19	19	20	21
15	7	8	9	10	19	20	21	22
16	8	8	9	10	20	20	22	22
17	8	9	10	10	20	21	22	22
18	8	9	10	11	20	21	22	23
19	9	9	10	11	21	22	23	23
20	9	10	10	11	21	22	23	23

TABLE II (*Concluded*)

$n = 14$

n	$u_{.005}$	$u_{.01}$	$u_{.025}$	$u_{.05}$	$u_{.95}$	$u_{.975}$	$u_{.99}$	$u_{.995}$
14	7	8	9	10	19	20	21	22
15	8	8	9	10	20	21	22	22
16	8	9	10	11	20	21	22	23
17	8	9	10	11	21	22	23	23
18	9	9	10	11	21	22	23	24
19	9	10	11	12	22	22	23	24
20	9	10	11	12	22	23	24	24

$n = 15$

n	$u_{.005}$	$u_{.01}$	$u_{.025}$	$u_{.05}$	$u_{.95}$	$u_{.975}$	$u_{.99}$	$u_{.995}$
15	8	9	10	11	20	21	22	23
16	9	9	10	11	21	22	23	23
17	9	10	11	11	21	22	23	24
18	9	10	11	12	22	23	24	24
19	10	10	11	12	22	23	24	25
20	10	11	12	12	23	24	25	25

$n = 16$

n	$u_{.005}$	$u_{.01}$	$u_{.025}$	$u_{.05}$	$u_{.95}$	$u_{.975}$	$u_{.99}$	$u_{.995}$
16	9	10	11	11	22	22	23	24
17	9	10	11	12	22	23	24	25
18	10	10	11	12	23	24	25	25
19	10	11	12	13	23	24	25	26
20	10	11	12	13	24	24	25	26

$n = 17$

n	$u_{.005}$	$u_{.01}$	$u_{.025}$	$u_{.05}$	$u_{.95}$	$u_{.975}$	$u_{.99}$	$u_{.995}$
17	10	10	11	12	23	24	25	25
18	10	11	12	13	23	24	25	26
19	10	11	12	13	24	25	25	26
20	11	11	13	13	24	25	26	27

$n = 18$

n	$u_{.005}$	$u_{.01}$	$u_{.025}$	$u_{.05}$	$u_{.95}$	$u_{.975}$	$u_{.99}$	$u_{.995}$
18	11	11	12	13	24	25	26	26
19	11	12	13	14	24	25	26	27
20	11	12	13	14	25	26	27	28

$n = 19$

n	$u_{.005}$	$u_{.01}$	$u_{.025}$	$u_{.05}$	$u_{.95}$	$u_{.975}$	$u_{.99}$	$u_{.995}$
19	11	12	13	14	25	26	27	28
20	12	12	13	14	26	26	28	28

$n = 20$

n	$u_{.005}$	$u_{.01}$	$u_{.025}$	$u_{.05}$	$u_{.95}$	$u_{.975}$	$u_{.99}$	$u_{.995}$
19								
20	12	13	14	15	26	27	28	29

TABLE III

Significance Levels of u

$$P\{u \leq u_\varepsilon\} = \varepsilon$$

For definition of u_ε, see note to Table II.

n=n	$u_{.005}$	$u_{.01}$	$u_{.025}$	$u_{.05}$	$u_{.95}$	$u_{.975}$	$u_{.99}$	$u_{.995}$
10	4	5	6	6	15	15	16	17
11	5	6	7	7	16	16	17	18
12	6	6	7	8	17	18	19	19
13	7	7	8	9	18	19	20	20
14	7	8	9	10	19	20	21	22
15	8	9	10	11	20	21	22	23
16	9	10	11	11	22	22	23	24
17	10	10	11	12	23	24	25	25
18	10	11	12	13	24	25	26	27
19	11	12	13	14	25	26	27	28
20	12	13	14	15	26	27	28	29
21	13	14	15	16	27	28	29	30
22	14	14	16	17	28	29	31	31
23	14	15	16	17	30	31	32	33
24	15	16	17	18	31	32	33	34
25	16	17	18	19	32	33	34	35
26	17	18	19	20	33	34	35	36
27	18	19	20	21	34	35	36	37
28	18	19	21	22	36	36	38	39
29	19	20	22	23	36	37	39	40
30	20	21	22	24	37	39	40	41
31	21	22	23	25	38	40	41	42
32	22	23	24	26	40	41	42	43
33	23	24	25	26	41	42	43	44
34	23	24	26	27	42	43	44	45
35	24	25	27	28	43	44	45	47
36	25	26	28	29	44	45	47	48
37	26	27	28	30	45	47	48	49
38	27	28	30	31	46	48	49	50
39	28	29	30	32	47	49	50	51
40	29	30	31	33	48	50	51	52
41	29	31	32	34	49	51	52	54
42	30	31	33	35	50	52	54	55
43	31	32	34	35	52	53	55	56
44	32	33	35	36	53	54	56	57
45	33	34	36	37	54	55	57	58
46	34	35	37	38	55	56	58	59
47	35	36	38	39	56	57	59	60
48	35	37	38	40	57	59	60	62
49	36	38	39	41	58	60	61	63
50	37	38	40	42	59	61	63	64
51	38	39	41	43	60	62	64	65
52	39	40	42	44	61	63	65	66
53	40	41	43	45	62	64	66	67
54	41	42	44	45	64	65	67	68
55	42	43	45	46	65	66	68	69

n=n	$u_{.005}$	$u_{.01}$	$u_{.025}$	$u_{.05}$	$u_{.95}$	$u_{.975}$	$u_{.99}$	$u_{.995}$
56	42	44	46	47	66	67	69	71
57	43	45	47	48	67	68	70	72
58	44	46	47	49	68	70	72	73
59	45	46	48	50	69	71	73	74
60	46	47	49	51	70	72	74	75
61	47	48	50	52	71	73	75	76
62	48	49	51	53	72	74	76	77
63	49	50	52	54	73	75	77	78
64	49	51	53	55	74	76	78	80
65	50	52	54	56	75	77	79	81
66	51	53	55	57	76	78	80	82
67	52	54	56	58	77	79	81	83
68	53	54	57	58	79	80	83	84
69	54	55	58	59	80	81	84	85
70	55	56	58	60	81	83	85	86
71	56	57	59	61	82	84	86	87
72	57	58	60	62	83	85	87	88
73	57	59	61	63	84	86	88	90
74	58	60	62	64	85	87	89	91
75	59	61	63	65	86	88	90	92
76	60	62	64	66	87	89	91	93
77	61	63	65	67	88	90	92	94
78	62	64	66	68	89	91	93	95
79	63	64	67	69	90	92	95	96
80	64	65	68	70	91	93	96	97
81	65	66	69	71	92	94	97	98
82	66	67	69	72	94	96	98	99
83	66	68	70	72	95	97	99	101
84	67	69	71	73	96	98	100	102
85	68	70	72	74	97	99	101	103
86	69	71	73	75	98	100	102	104
87	70	72	74	76	99	101	103	105
88	71	73	75	77	100	102	104	106
89	72	74	76	78	101	103	105	107
90	73	74	77	79	102	104	107	108
91	74	75	78	80	103	105	108	109
92	75	76	79	81	104	106	109	110
93	75	77	80	82	105	107	110	112
94	76	78	81	83	106	108	111	113
95	77	79	82	84	107	109	112	114
96	78	80	82	85	108	111	113	115
97	79	81	83	86	109	112	114	116
98	80	82	84	87	110	113	115	117
99	81	83	85	87	112	114	116	118
100	82	84	86	88	113	115	117	119

NOTES

This section is devoted to brief research and expository articles, and notes on methodology

A NOTE ON THE BEST LINEAR ESTIMATE

By Allen T. Craig

University of Iowa

1. Introduction. Let the chance variable x be subject to the distribution function $D(x)$ and as usual let $E[g(x)]$ denote the mathematical expectation of the function $g(x)$. If $x_1, x_2, \cdots x_n$ constitute a sample of n independent values of x, the function $y = c_1x_1 + c_2x_2 + \cdots + c_nx_n$ is frequently called the best linear estimate of $E(x)$ when the c's are so chosen that $E(y) = E(x)$, and $E[y - E(x)]^2 = \sigma_y^2$ is a minimum. It is the purpose of this note to give an example of an estimate y, best in the sense defined, yet such that, y' being another estimate,

$$Pr[E(x) - \delta \leq y \leq E(x) + \delta] \leq Pr[E(x) - \delta \leq y' \leq E(x) + \delta],$$

for every $\delta > 0$.

2. The rectangular distribution. Consider $D(x) = 1/a$, $0 \leq x \leq a$, and let the n items of each sample be arranged in ascending order of magnitude so that $x_1 \leq x_2 \leq \cdots \leq x_n$, $n \geq 2$. The generating function $G(t)$ of the moments of the distribution of $y = c_1x_1 + c_2x_2 + \cdots + c_nx_n$ is

$$G(t) = E(e^{ty}) = \frac{n!}{a^n}\int_0^a\int_0^{x_n}\int_0^{x_{n-1}} \cdots \int_0^{x_2} e^{t(c_1x_1 + \cdots + c_nx_n)}\,dx_1\,dx_2 \cdots dx_n.$$

Thus

$$E(y) = G'(0) = \frac{a}{n+1}[c_1 + 2c_2 + 3c_3 + \cdots + nc_n],$$

and

$$\begin{aligned} E(y^2) = G''(0) = \frac{a^2}{(n+1)(n+2)}[&1\cdot 2c_1^2 + 2\cdot 3c_2^2 + \cdots + n(n+1)c_n^2 \\ &+ 2\{1\cdot 3c_1c_2 + 1\cdot 4c_1c_3 + \cdots + 1\cdot(n+1)c_1c_n \\ &+ 2\cdot 4c_2c_3 + \cdots + 2(n+1)c_2c_n \\ &\qquad\vdots \\ &+ (n-1)(n+1)c_{n-1}c_n\}]. \end{aligned}$$

From $E(y) = E(x) = a/2$, we have

$$c_1 = \tfrac{1}{2}(n+1) - 2c_2 - \cdots - nc_n.$$

Thus $\sigma_y^2 = G''(0) - a^2/4$ with c_1 in $G''(0)$ replaced by $\frac{1}{2}(n-1) - 2c_2 - \cdots - nc_n$. From $\dfrac{\partial \sigma_y^2}{\partial c_j} = 0$, $j = 2, 3, \cdots, n$, we obtain the following system of $n-1$ non-homogeneous linear equations in $n-1$ unknowns

$$\begin{aligned}
4c_2 + 6c_3 + \cdots + 2nc_n &= n + 1 \\
6c_2 + 12c_3 + \cdots + 4nc_n &= 2(n+1) \\
8c_2 + 16c_3 + \cdots + 6nc_n &= 3(n+1) \\
&\vdots \\
2nc_2 + 4nc_3 + \cdots + 2n(n-1)c_n &= (n-1)(n+1).
\end{aligned}$$

Since the determinant of the coefficients is not zero, the solution $c_2 = c_3 = \cdots = c_{n-1} = 0$, $c_n = (n+1)/2n$, is unique. Further, we see that $c_1 = 0$ so the best linear estimate of the mean of the rectangular population is $y = (n+1)x_n/2n$, where x_n is the largest item in the sample.

The distribution function of y is readily found to be

$$D(y) = n\left[\frac{2n}{a(n+1)}\right]^n y^{n-1}, \qquad 0 \le y \le \frac{n+1}{2n}a.$$

From this, it follows that $\sigma_y^2 = \dfrac{a^2}{4n(n+2)}$.

It has long been known[1] that the sampling distribution of the statistic $\omega = \frac{1}{2}(x_1 + x_n)$, where x_1 and x_n are respectively the smallest and largest items in samples of size n from a rectangular population, has a smaller variance than does that of the arithmetic mean $\bar{x}$ of all n items. The distribution function of ω is

$$\begin{aligned}
D(\omega) &= \frac{2^{n-1}n\omega^{n-1}}{a^n}, && 0 \le \omega \le \tfrac{1}{2}a, \\
&= \frac{2^{n-1}n}{a^n}(a-\omega)^{n-1}, && \tfrac{1}{2}a \le \omega < a,
\end{aligned}$$

so that $E(\omega) = \frac{1}{2}a$ and $\sigma_\omega^2 = \dfrac{a^2}{2(n+1)(n+2)}$. Thus $\sigma_y^2 = \frac{1}{2}\sigma_\omega^2$, approximately. Yet Pittman has recently proved that for every $\delta > 0$, $Pr[E(x) - \delta \le \omega \le E(x) + \delta]$ exceeds the probability that any other estimate, including y, will fall in this interval of length 2δ about the mean $a/2$.

If we write $u = \dfrac{y - a/2}{\sigma_y}$ and $v = \dfrac{\omega - a/2}{\sigma_\omega}$, then the limits of $D(u)$ and $D(v)$ as n approaches infinity are respectively e^{u-1}, $-\infty \le u \le 1$, and $\dfrac{1}{\sqrt{2}}e^{-\sqrt{2}|v|}$,

[1] R. A. Fisher, "Theoretical foundations of mathematical statistics," *Phil. Trans. Roy. Soc. London*, Series A, Vol. 222 (1921), pp. 309–368.

$-\infty \leq v \leq \infty$. Thus neither y nor ω has an asymptotic normal distribution. It is, of course, this fact which makes the criterion of minimum variance illusory.

3. Other polynomial distribution functions. Let repeated samples of n independent values of x be drawn from a population characterized by $D(x) = \frac{k+1}{a^{k+1}}x^k$, $0 \leq x \leq a$, and k a positive integer or zero. It can be shown that the best linear estimate of the mean of the population is $y = \frac{(k+1)n+1}{n(k+2)}x_n$, where as before x_n is the largest item of the sample. The sampling distribution of y is easily obtained. It follows that

$$\sigma_y^2 = \frac{(k+1)a^2}{(k+2)^2[(k+1)n^2+2n]} = \frac{k+3}{n(k+1)+2}\sigma_{\bar{x}}^2,$$

where as usual $\bar{x}$ is the arithmetic mean of the sample. Again, if we write $u = \left(y - \frac{k+1}{k+2}a\right)\Big/\sigma_y$, the limit of the distribution of u as n approaches infinity is, as before, e^{u-1}, $-\infty \leq u \leq 1$.

A NOTE ON TOLERANCE LIMITS

By Edward Paulson[1]

Columbia University

Among various statistical problems arising in the process of controlling quality in mass production, a rather important one appears to be the determination of tolerance limits when the variability of the product is known to be due to random factors. This problem was recently treated in a pioneer article by Wilks. This note will point out a relationship between tolerance limits and confidence limits (used in the sense of Neyman), and will use this concept to establish tolerance limits when the product is described by two qualities, the measurements on which are assumed to have a bivariate normal distribution.

For the case of a single variate, the problem of finding tolerance limits as stated by Wilks is to find a sample size n, and two functions $L_1(x_1 \cdots x_n)$ and $L_2(x_1x_2 \cdots x_n)$ so that if $P = \int_{L_1}^{L_2} f(x)\,dx$ denotes the conditional probability of a future observation falling between the random variates L_2 and L_1, then

$$E(P) = \alpha, \quad \text{and Prob.} \quad [\alpha - \Delta_1 \leq P \leq \alpha + \Delta_2] \geq \beta.$$

The relationship between confidence limits and tolerance limits will arise if confidence limits are determined, not for a parameter of the distribution, but for

[1] Work done under a grant-in-aid from the Carnegie Corporation of New York.

a future random observation (or for some function of the observations in a future independent sample). This is based on the following simple lemma: *If confidence limits* $U_1(x_1 \cdots x_n)$ *and* $U_2(x_1 \cdots x_n)$ *on a probability level* $= \alpha_0$ *are determined for* g, *a function of a future sample of* k *observations, and* $P = \int_{U_1}^{U_2} \psi(g)\,dg$, *then* $E(P) = \alpha_0$ For let $\psi(g)\,dg$ and $\varphi(U_1, U_2)\,dU_1\,dU_2$ denote the distribution of g and U_1, U_2 respectively, then by the definition of expected value

$$E(P) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left[\int_{U_1}^{U_2} \psi(g)\,dg \right] \varphi(U_1, U_2)\,dU_1\,dU_2 .$$

This triple integral is however exactly the probability that g will lie between U_1 and U_2, which by the nature of confidence limits must equal α_0, which proves the lemma In a similar manner it follows that if on the basis of a given sample an l dimensional confidence region is found for statistics $g_1, g_2, \cdots g_l$ derived from a future sample, and if P denotes the probability that $g_1 \cdots g_l$ all fall in the confidence region, then $E(P)$ in repeated sampling equals α To establish tolerance limits, it is necessary in addition to $E(P)$ to also know the distribution of P, or at least σ_P^2, so the distribution of P can be approximated.

It appears, at least on an intuitive basis, that the "best" confidence interval can be used to determine the shape of the "most efficient" tolerance limits; this intuitive notion will gain additional support from the character of the tolerance region which will now be derived for an observation (x, y) from a distribution with probability density $f(x, y)$, where

$$f(x, y) = \frac{\exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{x - m_x}{\sigma_x} \right)^2 - 2\rho \left(\frac{x - m_x}{\sigma_x} \right) \left(\frac{y - m_y}{\sigma_y} \right) + \left(\frac{y - m_y}{\sigma_y} \right)^2 \right] \right\}}{2\pi \sigma_x \sigma_y \sqrt{1 - \rho^2}}$$

Suppose we have 2 independent samples

$$[(x_1, y_1)(x_2, y_2) \cdots (x_n, y_n)] \quad \text{and} \quad [(x, y)]$$

both from $f(x, y)$ Then it is known that

$$T^2 = \left(\frac{n}{n+1} \right) \frac{1}{1 - r^2} \left\{ \left(\frac{\bar{x} - x}{s_x} \right)^2 - \frac{2r}{s_x s_y} (\bar{x} - x)(\bar{y} - y) + \frac{(\bar{y} - y)^2}{s^2 y} \right\}$$

where $\bar{x} = \sum_{i=1}^{n} x_i/n$, $s_x^2 = \sum_{1}^{n} (x_i - \bar{x})^2/(n - 1)$, etc, has the distribution of Hotelling's Generalized Student Ratio [2]. A confidence region for a future observation (x, y) on the basis of a sample of n on a level of significance $= \alpha$ will be given by the elliptic region $T^2 \leq T_\alpha^2$ (in the x, y plane), where $T_\alpha^2 = 2(n - 1) F_0/(n - 2)$, where F_0 is the value of the F distribution (with $n_1 = 2$ and $n_2 = n - 2$ degrees of freedom) which is exceeded with probability $= 1 - \alpha$.

If P denotes the probability of a future observation falling in this ellipse, then $P = \iint_{T^2 \leq T_\alpha^2} f(x, y)\,dx\,dy$. By utilizing the fact [2] that T^2 is invariant under linear

transformations, it is not difficult to see that the distribution of P will not involve any unknown parameters, so its distribution can be calculated under the assumption $m_x = m_y = \rho = 0$, $\sigma_x = \sigma_y = 1$. Then

$$P = F(\bar{x}, \bar{y}, s_x, s_y, r) = \iint_{I^2 + J^2 \le T_\alpha^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}\, dx\, dy.$$

We know that $E(P) = \alpha$, and we will now calculate the variance of P by expanding P in a Taylor Series (to terms of the first order) about the point $\bar{x} = 0$, $\bar{y} = 0$, $r = 0$, $s_x = 1$, $s_y = 1$. P can clearly be put in the form

$$P = \frac{1}{2\pi} \int_{x - T_\alpha \sqrt{\frac{n+1}{n}} s_x}^{x + T_\alpha \sqrt{\frac{n+1}{n}} s_x} e^{-\frac{1}{2}x^2}\, dx \int_{y + r s_y \left(\frac{x-\bar{x}}{s_x}\right) - s_y \sqrt{(1-r^2)\left[T_\alpha^2\left(\frac{n+1}{n}\right) - \left(\frac{x-\bar{x}}{s_x}\right)^2\right]}}^{\bar{y} + r s_y \left(\frac{x-\bar{x}}{s_x}\right) + s_y \sqrt{(1-r^2)\left[T_\alpha^2\left(\frac{n+1}{n}\right) - \left(\frac{x-\bar{x}}{s_x}\right)^2\right]}} e^{-\frac{1}{2}y^2}\, dy$$

Taking derivatives and evaluating about the population values

$$\left[\frac{\partial P}{\partial \bar{x}}\right] = \left[\frac{\partial P}{\partial \bar{y}}\right] = \left[\frac{\partial P}{\partial r}\right] = 0,$$

$$\left[\frac{\partial P}{\partial s_x}\right] = \frac{e^{-\frac{1}{2}T_\alpha^2\left(\frac{n+1}{n}\right)}}{\pi} \int_{-T_\alpha \sqrt{\frac{n+1}{n}}}^{T_\alpha \sqrt{\frac{n+1}{n}}} x^2\, dx \sqrt{T_\alpha^2\left(\frac{n+1}{n}\right) - x^2}$$

$$= \tfrac{1}{2} e^{-\frac{1}{2}T_\alpha^2\left(\frac{n+1}{n}\right)} T_\alpha^2\left(\frac{n+1}{n}\right)$$

$$\left[\frac{\partial P}{\partial s_y}\right] = \frac{e^{-\frac{1}{2}T_\alpha^2\left(\frac{n+1}{n}\right)}}{\pi} \int_{-T_\alpha \sqrt{\frac{n+1}{n}}}^{T_\alpha \sqrt{\frac{n+1}{n}}} \sqrt{T_\alpha^2\left(\frac{n+1}{n}\right) - x^2}\, dx$$

$$= e^{-\frac{1}{2}T_\alpha^2\left(\frac{n+1}{n}\right)} \tfrac{1}{2} T_\alpha^2\left(\frac{n+1}{n}\right).$$

So $$\delta P = e^{-\frac{1}{2}T_\alpha^2\left(\frac{n+1}{n}\right)} \tfrac{1}{2} T_\alpha^2\left(\frac{n+1}{n}\right)[\delta s_x + \delta s_y],$$

and to terms of $0\left(\frac{1}{n}\right)$:

$$\sigma_P^2 = \frac{T_\alpha^4 e^{-T_\alpha^2}}{4n}.$$

Since for ordinary values of $\alpha (\alpha = .95$ or $.99)$ the distribution of P seems to approach normality very slowly, we will follow a suggestion of Wilks and suppose that a fairly close approximation to the distribution of P will be given by

$$(1) \qquad \frac{\Gamma(u+v)}{\Gamma(u)\Gamma(v)} P^{u-1}(1-P)^{v-1},$$

where $u = [\alpha^2(1 - \alpha) - \alpha\sigma_P^2]/\sigma_P^2$

$v = [\alpha(1 - \alpha)^2 - (1 - \alpha)\sigma_P^2]/\sigma_P^2$.

This distribution can now be used to establish tolerance limits. For example, it follows from (1) that for a sample size $n \geq 214$, and a tolerance region given by the ellipse $T^2 = 9.21$, then $E(P) = .99$ and the Prob.$\{.985 \leq P \leq .995\} \geq .992$.

Care must be taken in the use of these and similar results, for if the distribution is not a bivariate normal one, a large error may be introduced which will not be eliminated with increasing n; however the error will probably be small when a tolerance region is found for the means $\bar{x}$, $\bar{y}$ of a future sample of k observations ($k \geq 20$) as contrasted with a tolerance region for a single observation. An exact treatment of the case when the bivariate distribution is unknown has been given by Wald in the present issue of the *Annals of Mathematical Statistics*.

REFERENCES

[1] S. S. Wilks, "Determination of sample sizes for setting tolerance limits," *Annals of Math. Stat.*, Vol 12 (1941), pp. 91–96.

[2] Harold Hotelling, "A generalization of Student's ratio," *Annals of Math. Stat.*, Vol 2 (1931), pp 360–378.

A NEW APPROXIMATION TO THE LEVELS OF SIGNIFICANCE OF THE CHI-SQUARE DISTRIBUTION.

By Leo A. Aroian

Hunter College

Recent articles on the percentage points of the χ^2 distribution [1], [2], have directed my attention to a method proposed in my investigation of Fisher's z distribution [3], a method particularly useful and easily computed for n large. In addition, this method avoids interpolation. If $t = \dfrac{\chi^2 - n}{\sqrt{2n}}$, and $\alpha_3 = \sqrt{\dfrac{8}{n}}$, the measure of skewness for the χ^2 distribution, the following formulas give significance levels of t as quadratic functions of α_3, $t = a + b\alpha_3 + c\alpha_3^2$. The values of a, b, and c were found by the usual method of least squares, fitting each formula to the values of t [4] for $\alpha_3 = 0, \pm 0.1, \pm 0.2, \pm 0.3$, and ± 0.4. Then the value of a in each instance was adjusted to give the proper value of t when $\alpha_3 = 0$; e.g. the constant term by the method of least squares for the 1 per cent point is 2.32633 which we change to 2.32635. The range $|\alpha_3| \leq .4$ corresponds to $n \geq 50$, but the formulas are quite satisfactory for $n \geq 30$. Formulas for t when $|\alpha_3| > .4$ [3] are easily derived, but such results while more accurate in the range

TABLE I[1]

n \ P	.995	.99	95	90	75	50	25	10	.05	01	005
30	13 744	14.925	18 491	20 604	24.486	29.338	34.793	40.246	43.767	50 914	53 713
	13.7997	14.9649	18 4960	20.6004	24.4764	29 3346	34.7987	40.2559	43.7754	50 9015	53.6883
	13 7867	14 9535	18.4926	20 5992	24 4776	29 3360	34.7998	40.2560	43.7729	50.8922	53 6720
40	20 669	22 139	26.5080	29.055	33.668	39.337	45.610	51.796	55.753	63 710	66.802
	20.7121	22 1703	26 5114	29 0514	33.6597	39 3346	45.6155	51.8048	55.7600	63 6961	66 776
	20.7065	22.1643	26.5093	29.0505	33.6603	39.3354	45.6160	51.8050	55 7585	63 6907	66 7659
50	27.957	29 685	34.7634	37.693	42.949	49 336	56 328	63.159	67.501	76.172	79.523
	27.9920	29.7096	34 7656	37.6894	42.9418	49.3346	56.3333	63 1669	67 5057	76 1568	79.496
	27.9907	29.7067	34.7642	37 6886	42.9421	49 3349	56.3336	63.1671	67.5048	76 1539	79 4900
75	47.178	49 457	56 0538	59 799	66 422	74 335	82 853	91 055	96 2135	106 408	110.313
	47.2021	49.4741	56 0546	59 7951	66.4169	74 3346	82.8581	91.0611	96.2168	106 392	110 286
	47.2059	49 4748	56.0540	59 7944	66.4167	74 3343	82.8582	91.062	96 2160	106 393	110 286
100	67.303	70.049	77 9294	82 362	90.138	99 335	109.137	118 493	124.340	135.820	140.193
	67.3210	70 0620	77.9296	82 3586	90 1336	99 3346	109.1416	118 498	124.342	135.804	140.166
	67.3276	70.0648	77.9295	82 3581	90.1332	99.3341	109.141	118.498	124.342	135.807	140 169

[1] First value in cell by method of Wilson and Hilferty.
Second value in cell by (1) or (2)
Third value is correct result.

TABLE II[2]

P	.9999	999	975	80	70	60	40	30	20	025	001	0001
$n = 30$	9.33	11.62	16 7962	23.3631	25.5064	27.4402	31.3144	33.5290	36 2494	46 9821	59 73	67.68
	9.226	11.568	16.7908	23.364	25.508	27.4436	31.3183	33.530	36.250	46.9792	59.683	67.590

[2] First value by (1) or (2).
Second value correct result.

$30 \leqq n < 50$ would be considerably less accurate in the region $n \geqq 50$. After t is calculated, $\chi^2 = n + \sqrt{2n}\,t$ The formulas are:

(1)
$$t_{50\%} = -.16636\alpha_3$$
$$t_{40\%} = .25335 - .15567\alpha_3 - .012276\alpha_3^2$$
$$t_{30\%} = .52440 - .12058\alpha_3 - .021541\alpha_3^2$$
$$t_{25\%} = .67449 - .090613\alpha_3 - .030693\alpha_3^2$$
$$t_{20\%} = .84162 - .048433\alpha_3 - .036788\alpha_3^2$$
$$t_{10\%} = 1.28155 + .107033\alpha_3 - .04797\alpha_3^2$$
$$t_{5\%} = 1.64485 + .28392\alpha_3 - .04902\alpha_3^2$$
$$t_{2\,5\%} = 1.95997 + 47228\alpha_3 - .04304\alpha_3^2$$
$$t_{1\%} = 2\ 32635 + .73330\alpha_3 - .024957\alpha_3^2$$
$$t_{.5\%} = 2.5758 + .93600\alpha_3 - .00377\alpha_3^2$$
$$t_{.1\%} = 3.0903 + 1.4190\alpha_3 + .05667\alpha_3^2$$
$$t_{.01\%} = 3.7200 + 2.1260\alpha_3 + .17449\alpha_3^2$$

The maximum error for t in the range $|\alpha_3| \leqq .4$, is 2 in the fourth significant figure, 1 in the fourth significant figure, 6 in fifth, 3 in fifth, 3 in fifth, 1 in fifth, 1 in fifth, 3 in fifth, 4 in fifth, 4 in fifth, 4 in fifth and 4 in fourth significant figures respectively for the .01%, .1%, 5%, 1%, 2.5%, 5%, 10%, 20%, 25%, 30%, 40%, and 50% points respectively The error increases outside the indicated range. In addition

(2)
$$t_{99.99\%} = -3.7200 + 2.1260\alpha_3 - .17449\alpha_3^2$$
$$t_{99.9\%} = -3.0903 + 1.4190\alpha_3 - .05667\alpha_3^2$$

and similarly for other percentage points. These are obtained from (1) by replacing α_3 by $-\alpha_3$ and t by $-t$.

We compare results obtained by these methods against those of Wilson and Hilferty [2]. In all cases except at the 95% level the method here proposed is superior. Table I compares the two methods. It was copied from [2] except for the corrections in the Wilson and Hilferty method for the 95% level and in the accurate value for χ^2 at the 5% level for $n = 75$, 96.2160 in place of 96. 11. Table II gives comparisons for other levels when $n = 30$

REFERENCES

[1] C. M. Thompson, "Table of percentage points of the χ^2 distribution," *Biometrika*, Vol. 32, Part 2

[2] M. Merrington, "Numerical approximations to the percentage points of the χ^2 distribution," *Biometrika*, Vol 32, Part 2.

[3] L. A Aroian, "A study of R A Fisher's z distribution and the related F distribution," *Annals of Math. Stat.*, Vol. 12 (1941).

[4] L R. Salvosa, "Tables of Pearson's Type III function," *Annals of Math. Stat.*, Vol. 1.

NEWS AND NOTICES

Readers are invited to submit to the Secretary of the Institute news items of general interest

Personal Items

Associate Professor H. P. Evans of the Mathematics Department of the University of Wisconsin has been promoted to a professorship.

Assistant Professor Willy Feller of the Mathematics Department of Brown University has been promoted to an associate professorship.

Dr. Carl F. Kossack of the Mathematics Department of the University of Oregon has been promoted to an assistant professorship.

Dr. Eugene Lukacs has been appointed to an assistant professorship in the Mathematics Department of Illinois College

Professor E. B. Mode has been made chairman of the Mathematics Department of Boston University.

Mr. Charles R. Mummery has been made Product Quality Engineer at the Scioto Ordinance Plant of the U. S. Rubber Company.

Professor H. L. Rietz has retired after twenty-five years of service as Head of the Mathematics Department of the University of Iowa.

The Foundation for the Study of Cycles has announced that a medal will be awarded to the individual making the most significant contribution to cycle research during 1943. Communications should be addressed to: Professor Ellsworth Huntington, Yale University, New Haven, Connecticut.

Obituary

Professor Edward L. Dodd of the Mathematics Department of the University of Texas died on January 9, 1943 at the age of sixty-seven years. He was a charter member of the Institute. He was elected as one of the Vice-Presidents of the Institute for 1943. His contributions to mathematical statistics consist of numerous research papers on probability, on general mean functions of statistical variables, and on statistical theory of periodicities.

Stanford Courses in Statistical Methods of Quality Control

A novel procedure in adult education, and particularly in statistical education, took place last summer at Stanford University, when courses in the Shewhart statistical methods of quality control were offered in short intensive courses. There were two courses, one on the campus at Stanford University, July 17–26, and the other in Los Angeles, September 20–27. The first course covered ten full days, and the second eight. Both courses ran eight hours per day, Saturdays and Sundays included.

The features of the course may be described by the following points:

1. The courses were short, thus making it possible for men in industry to attend
2. The number of hours' instruction was sufficient to cover the field adequately.
3. The instruction covered a wide range of points of view.
4. The students were picked delegates sent by industry.
5. The courses are being followed up with monthly meetings in Los Angeles and San Francisco.

Intensive courses of this character were first suggested by Dr. W. Edwards Deming in April of 1942, while he was temporarily detailed to the office of the Chief of Ordnance in the War Department, and the first course actually commenced just three months later. By giving the course to men already in industry, the yield obtained was manyfold higher than can be expected from a regular college course. Contributions and reports made by the delegates subsequent to the courses supply abundant foundation for this statement.

West Coast industry and the Army and Navy ordnance districts sent 32 delegates to the West Coast, and 31 to the second. Through the efforts of Professor Eugene Grant of Stanford, industry and the Army and Navy were persuaded to send some of their most valued officials. The instruction was organized by Professor Holbrook Working. Both he and Professor Grant took an active part in the instruction, which was supplemented in both courses by Dr. W. Edwards Deming as an exponent of government and industrial sampling. In the first course, Mr. Charles R Mummery of the Hoover Company served as an instructor from the viewpoint of industry. In the second course (the one in Los Angeles), Mr. Ralph E. Wareham of the General Electric Company occupied the industrial corner of the square of instruction. The expense of the instructors was paid out of ESMWT funds (Office of Education). Montly follow-up courses in San Francisco and Los Angeles, under the direction of Professors Working and Grant, supply the necessary power for maintaining momentum, and for gathering the men together for directed study and consultation.

The demand for men trained in this line far exceeds the supply, and there are movements afoot to provide similar courses in a number of industrial cities. Three-day courses in a dozen or more key ordnance cities were held last fall by the Ordnance Department The lecturers were Messrs. G. D. Edwards and Harold F Dodge of the Bell Telephone Laboratories, and Mr. G. Rupert Gause of the Aberdeen Proving Ground, now with the Army Ordnance in Washington. These courses and the Stanford courses alleviated the situation considerably, but further instruction is needed

Junior Membership in the Institute

At the annual election for 1942 which was held by mail ballot because of the postponement of the Annual Meeting, constitutional amendments were approved which created a new grade of membership in the Institute, known as Junior

Membership It is hoped that this provision for Junior Membership will stimulate interest in mathematical statistics at the advanced undergraduate level in colleges and universities.

The Board of Directors have approved the following rules governing Junior Membership:

1. Any undergraduate student of a collegiate institution is eligible for election as a Junior member of the Institute of Mathematical Statistics provided that he or she is sponsored by a member of the Institute
2. The annual dues ($2.50) must be submited with the application
3. Annual membership shall coincide with the calendar year and the Junior Member shall receive a complete volume of the *Annals of Mathematical Statistics* for the year in which he or she is elected.
4. Junior Membership shall be limited to a term of two years, but a Junior Member may apply for transfer to ordinary membership at the beginning of his second year.

For the convenience of any Institute member who may wish to sponsor a Junior Member an application blank is provided at the back of this issue of the *Annals*. Additional blanks may be obtained from the Secretary of the Institute.

Announcement of May Meeting in New York

There will be a joint meeting between the Institute and the American Society of Mechanical Engineers on Saturday, May 29, 1943, at the Engineering Societies Building, 29 West 39th Street, N. Y.

The meeting will consist of two sessions on industrial applications of mathematical statistics. The topics are as follows:

Morning Session, 10 A.M.

Chairman: Harold Hotelling

1 J. Wolfowitz, *On the Theory of Runs with some Applications to Quality Control.*

2 Churchill Eisenhart, *On the Presentation of Data as Evidence.*

Afternoon Session, 2 P.M.

Chairman: W. A. Shewhart

1. H. F. Dodge, *A Sampling Inspection Plan for Continuous Production.*
2. L. C. Young, *Tolerances and Product Acceptability.*

ANNUAL REPORT OF THE PRESIDENT OF THE INSTITUTE

Ordinarily at the business meeting and at the luncheon customarily held as part of the annual meeting of the Institute, the President has the opportunity to make public acknowledgement to those individuals, aside from the officers, who served the Institute during the year, and to have his say concerning past progress and future plans. This year it appears that the pages of the *Annals* must be used for this purpose.

As the result of proposals made and approved at our last regular annual meeting in New York City, a larger number of special committees than usual were appointed for 1942 and thus more members than before were specifically asked to participate in the affairs of the Institute. The Institute is much indebted to these individuals for the way in which they responded.

Professors A. T. Craig, Harold Hotelling, and S. S. Wilks, Chairman, constituted a committee to study the Board of Directors of the Institute and the formal connection between the Institute and its journal, *The Annals of Mathematical Statistics*. Their recommendations were incorporated in amendments to the constitution and by-laws recently approved by the Institute. The Board was increased in size and given greater continuity by including in it the two previous presidents, and the editor of the *Annals*, ex officio, and by increasing the term of the Secretary-Treasurer to three years.

The new class of junior memberships is the result of a study of this question by a committee composed of Professors J. H. Bushey, Boyd Harshbarger, and G. W. Snedecor, Chairman. Regulations, since approved by the Board, under which local chapters of the Institute may be formed, were drawn up by a committee consisting of Dr. C. F. Kossack, Professor H. D. Larsen, and Professor B. H. Camp, Chairman.

Dr. L. A. Aroian, Dr. J. F. Daly, Mr. H. F. Dodge, and Professor W. D. Baten, Chairman, as a committee agreed to assist the Institute by endeavoring to bring the *Annals of Mathematical Statistics* to the favorable attention of municipal, industrial, and college libraries which had not been subscribers to it. As a result of their fine work, a good number of domestic libraries has been added to our subscription list, thus serving to counterbalance our losses abroad.

The Program Committees for the year consisted of Professor Churchill Eisenhart and Mr. E. C. Molina, Chairman, for the September meeting in Poughkeepsie, New York, and of Professor P. S. Dwyer and Dr. W. E. Deming, chairman, for the projected Cleveland meeting. The Institute is always much indebted to those who do the work of arranging its programs for meetings, but this year we owe Dr. Deming a special acknowledgement, who prepared an excellent program for Cleveland, then one for a New York meeting under extremely short notice when the Cleveland meeting was cancelled, and then had that meeting also cancelled. Dr. W. R. Van Voorhis acted as our representative on the Committee on Local Arrangements for the meeting planned for Cleveland.

The membership committee appointed for 1942 was made up of Dr. W. E. Deming, Professor E. L. Dodd, and Professor A. T. Craig, chairman. After Professor Craig took up his commission in the Navy, Professor B. H. Camp agreed to take his place on this committee.

For some years Professor A. T. Craig has generously acted as custodian of our files of back numbers of the *Annals*. This service to the Institute has been taken over by Professor L. A. Knowler, who has been of much assistance. Dr. W. E. Blanche did a considerable amount of work in connection with finding advertisers for the *Annals*.

Our annual meetings had been increasingly successful in recent years and it was a real sacrifice for the Institute to forego the one planned for 1942. Though the war is demonstrating in still more ways and places the importance of sound statistical methods, for the present it imposes serious responsibilities on the friends of the Institute. A reading of the report of our faithful and efficient Secretary-Treasurer will amplify this statement. Of the present Board, Professors Olds, Wilks and Craig met in Pittsburgh January 23 and 24 to consider some of our problems. Though there seems no prospect of a national meeting in the coming year, it is hoped that some local meetings can be held and that in other ways we can keep up the activities of the Institute In particular there exists the opportunity of organizing local chapters of the Institute in the larger centers which would be particularly valuable now In industrial areas we may contribute to the war effort as well as promote an important aspect of mathematical statistics by endeavoring to be useful in the development and application of industrial statistics. It is clear that the Institute needs the loyal support of its membership now as much as ever before if it is to fulfill the functions for which it was founded.

CECIL C. CRAIG,
President.

December 31, 1942

ANNUAL REPORT OF THE SECRETARY-TREASURER OF THE INSTITUTE

On September 8–9 the Institute met at Vassar College in conjunction with the American Mathematical Society and the Mathematical Association of America. Mr. E. C. Molina and Professor Churchill Eisenhart were in charge of the program Fifty eight members of the Institute attended the meeting. The Annual Meeting, originally scheduled for Cleveland then transferred to New York City, was finally postponed at the request of the Office of Defense Transportation At the present time it seems that this meeting will have to be abandoned entirely and the Institute must be content with holding local meetings in some of the larger cities.

Because of the postponement of the Annual Meeting, the annual election was held by mail. The following officers were elected: Professor Cecil C. Craig, President, Professors Edward L. Dodd and Abraham Wald, Vice-Presidents; and Professor Edwin G. Olds, Secretary-Treasurer Nine amendments to the Constitution and six amendments to the By-Laws were proposed and accepted by a two-thirds majority of those voting. Professor K. L. Fetters acted as teller.

During the past year the Secretary has cooperated with industrial concerns and government agencies in locating statistically trained personnel to fill positions created by the emergency. Members of the Institute are requested to keep the Secretary informed regarding the availability of such personnel.

The death of one member of the Institute has been reported since the last

annual meeting—Dr. Robert Henderson, former Vice-President and Actuary of the Equitable Life Assurance Society

The following financial statement covers the period from December 10, 1941 to December 10, 1942 (the books and records of the Treasurer have been audited by Mr. George E Niver and found to be in agreement with the statement as submitted):

FINANCIAL STATEMENT

December 10, 1941, to December 10, 1942

RECEIPTS

BALANCE ON HAND, December 10, 1941		$1,561 54
DUES		2,203.15
SUBSCRIPTIONS		1,289.74
SALES OF BACK NUMBERS		1,393 65
CUMULATIVE INDEX		5 00
MISCELLANEOUS		28 50
Total Receipts		$6,571 58

EXPENDITURES

ANNALS OFFICE		
Editorial Expenses.		$127 08
WAVERLY PRESS		
Printing and Mailing Annals—4 issues		3,227.34
BACK NUMBERS OFFICE		
Purchase of back numbers from H C. Carver	$355 77	
Reprinting 300 copies of Vol V, No 4.	142.16	
		$497.93
LIBRARY COMMITTEE		16 77
SECRETARY-TREASURER'S OFFICE		
Printing and Supplies	$53.58	
Binding	30.00	
Postage	130.16	
Clerical Help	215.50	
		$438 24
PRINTING PROGRAMS FOR MEETINGS		101 11
MISCELLANEOUS		7 08
Total Expenditures		$4,416 45
BALANCE ON HAND, December 10, 1942		2,155 13
		$6,571.58

In comparison with the financial condition of the Institute at the end of 1941, the receipts from dues, subscriptions, and sales of back numbers have increased more than $800. This is mostly due to a large increase in the sales of back numbers and a net increase of fifty members The increase in expenditures of the Institute was accounted for by the increased cost in printing the *Annals*. This marks the beginning of a trend which seems likely to continue throughout the war.

It would seem over-optimistic to expect that the financial situation of the Institute would continue to show marked improvement in 1943. Present indications suggest that we shall be very fortunate to avoid a considerable deficit in operations. A large number of our foreign subscribers have not renewed and we face considerable difficulty in delivery of the *Annals* to those still in force. The large increase in the sales of back numbers was due to a rather successful effort to persuade domestic libraries to provide themselves with complete sets of back numbers while the issues were still available. The Institute faces an increase in operating expenses and an advance in the cost of producing the *Annals*. The full cooperation of all members is needed if we are to avoid a decrease in the work of the Institute during 1943.

EDWIN G. OLDS,
Secretary-Treasurer.

December 31, 1942.

On behalf of the Board of Directors of the Institute, I regret to announce the sudden death of Vice-President E. L. Dodd, on January 9, 1943, shortly after this report was written. Dr. W. E. Deming was appointed by the Board of Directors to fill the vacancy created by Vice-President Dodd's death.

E. G. O.

CONSTITUTION
OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

ARTICLE I

NAME AND PURPOSE

1. This organization shall be known as the Institute of Mathematical Statistics.
2. Its object shall be to promote the interests of mathematical statistics.

ARTICLE II

MEMBERSHIP

1. The membership of the Institute shall consist of Members, Junior Members, Fellows, Honorary Members, and Sustaining Members.

2. Voting members of the Institute shall be (a) the Fellows, and (b) all others, Junior Members excepted, who have been members for twenty-three months prior to the date of voting.

3. No person shall be a Junior Member of the Institute for more than a limited term as determined by the Committee on Membership and approved by the Board of Directors.

ARTICLE III

OFFICERS, BOARD OF DIRECTORS, AND COMMITTEE ON MEMBERSHIP

1. The Officers of the Institute shall be a President, two Vice-Presidents, and a Secretary-Treasurer. The terms of office of the President and Vice-Presidents shall be one year

and that of the Secretary-Treasurer three years. Elections shall be by majority ballot at Annual Meetings of the Institute. Voting may be in person or by mail.

(a) Exception. The first group of Officers shall be elected by a majority vote of the individuals present at the organization meeting, and shall serve until December 31, 1940.

2. The Board of Directors of the Institute shall consist of the Officers, the two preceding Presidents, and the Editor of the Official Journal of the Institute.

3. The Institute shall have a Committee on Membership composed of three Fellows. At their first meeting subsequent to the adoption of this Constitution, the Board of Directors shall elect three members as Fellows to serve as the Committee on Membership, one member of the Committee for a term of one year, another for a term of two years, and another for a term of three years. Thereafter the Board of Directors shall elect from among the Fellows one member annually at their first meeting after their election for a term of three years. The president shall designate one of the Vice-Presidents as Chairman of this Committee.

ARTICLE IV

Meetings

1. A meeting for the presentation and discussion of papers, for the election of Officers, and for the transaction of other business of the Institute shall be held annually at such time as the Board of Directors may designate. Additional meetings may be held from time to time by the Board of Directors and shall be called at any time by the President upon written request from ten Fellows. Notice of the time and place of meeting shall be given to the membership by the Secretary-Treasurer at least thirty days prior to the date set for the meeting. All meetings except executive sessions shall be open to the public. Only papers accepted by a Program Committee appointed by the President may be presented to the Institute.

2. The Board of Directors shall hold a meeting immediately after their election and again immediately before the expiration of their term. Other meetings of the Board may be held from time to time at the call of the President or any two members of the Board. Notice of each meeting of the Board, other than the two regular meetings, together with a statement of the business to be brought before the meeting, must be given to the members of the Board by the Secretary-Treasurer at least five days prior to the date set therefor. Should other business be passed upon, any member of the Board shall have the right to reopen the question at the next meeting.

3. The Committee on Membership shall hold a meeting immediately after the annual meeting of the Institute. Further meetings of the Committee may be held from time to time at the call of the Chairman or any member of the Committee provided notice of such call and the purpose of the meeting is given to the members of the Committee by the Secretary-Treasurer at least five days before the date set therefor. Should other business be passed upon, any member of the Committee shall have the right to reopen the question at the next meeting.

4. At a regularly convened meeting of the Board of Directors, four members shall constitute a quorum. At a regularly convened meeting of the Committee on Membership, two members shall constitute a quorum.

ARTICLE V

Publications

1. The *Annals of Mathematical Statistics* shall be the Official Journal for the Institute. The Editor of the *Annals of Mathematical Statistics* shall be a Fellow appointed by the

Board of Directors of the Institute. The term of office of the Editor may be terminated at the discretion of the Board of Directors

2 Other publications may be originated by the Board of Directors as occasion arises

ARTICLE VI

EXPULSION OR SUSPENSION

1 Except for non-payment of dues, no one shall be expelled or suspended except by action of the Board of Directors with not more than one negative vote.

ARTICLE VII

AMENDMENTS

1 This constitution may be amended by an affirmative two-thirds vote at any regularly convened meeting of the Institute provided notice of such proposed amendment shall have been sent to each voting member by the Secretary-Treasurer at least thirty days before the date of the meeting at which the proposal is to be acted upon. Voting may be in person or by mail.

BY-LAWS

ARTICLE I

DUTIES OF THE OFFICERS, THE EDITOR, BOARD OF DIRECTORS, AND COMMITTEE ON MEMBERSHIP

1. The President, or in his absence, one of the Vice-Presidents, or in the absence of the President and both Vice-Presidents, a Fellow selected by vote of the Fellows present, shall preside at the meetings of the Institute and of the Board of Directors. At meetings of the Institute, the presiding officer shall vote only in the case of a tie, but at meetings of the Board of Directors he may vote in all cases. At least three months before the date of the annual meeting, the President shall appoint a Nominating Committee of three members It shall be the duty of the Nominating Committee to make nominations for Officers to be elected at the annual meeting and the Secretary-Treasurer shall notify all voting members at least thirty days before the annual meeting Additional nominations may be submitted in writing, if signed by at least ten Fellows of the Institute, up to the time of the meeting.

2. The Secretary-Treasurer shall keep a full and accurate record of the proceedings at the meetings of the Institute and of the Board of Directors, send out calls for said meetings and, with the approval of the President and the Board, carry on the correspondence of the Institute Subject to the direction of the Board, he shall have charge of the archives and other tangible and intangible property of the Institute, and once a year he shall publish in the *Annals of Mathematical Statistics* a classified list of all Members and Fellows of the Institute. He shall send out calls for annual dues and acknowledge receipt of same; pay all bills approved by the President for expenditures authorized by the Board or the Institute; keep a detailed account of all receipts and expenditures, prepare a financial statement at the end of each year and present an abstract of the same at the annual meeting of the Institute after it has been audited by a Member or Fellow of the Institute appointed by the President as Auditor. The Auditor shall report to the President.

3. Subject to the direction of the Board, the Editor shall be charged with the responsibility for all editorial matters concerning the editing of the *Annals of Mathematical Statistics*. He shall, with the advice and consent of the Board, appoint an Editorial Committee

tee of not less than twelve members to co-operate with him; four for a period of five years, four for a period of three years, and the remaining members for a period of two years, appointments to be made annually as needed All appointments to the Editorial Committee shall terminate with the appointment of a new Editor. The Editor shall serve as editorial adviser in the publication of all scientific monographs and pamphlets authorized by the Board.

4. The Board of Directors shall have charge of the funds and of the affairs of the Institute, with the exception of those affairs specifically assigned to the President or to the Committee on Membership. The Board shall have authority to fill all vacancies ad interim, occurring among the Officers, Board of Directors, or in any of the Committees. The Board may appoint such other committees as may be required from time to time to carry on the affairs of the Institute.

5. The Committee on Membership shall prepare and make available through the Secretary-Treasurer an announcement indicating the qualifications requisite for the different grades of membership.

ARTICLE II

Dues

Members shall pay five dollars at the time of admission to membership and shall receive the full current volume of the Official Journal Thereafter, Members shall pay five dollars annual dues The annual dues of Junior Members shall be two dollars and fifty cents.

The annual dues of Fellows shall be five dollars The annual dues of Sustaining Members shall be fifty dollars Honorary Members shall be exempt from all dues.

(a) Exception. In the case that two Members of the Institute are husband and wife and they elect to receive between them only one copy of the Official Journal, the annual dues of each shall be three dollars and seventy-five cents.

2. Annual dues shall be payable on the first day of January of each year.

3. The annual dues of a Fellow, Member, or Junior Member include a subscription to the Official Journal The annual dues of a Sustaining Member include two subscriptions to the Official Journal.

4. It shall be the duty of the Secretary-Treasurer to notify by mail anyone whose dues may be six months in arrears, and to accompany such notice by a copy of this Article. If such person fail to pay such dues within three months from the date of mailing such notice, the Secretary-Treasurer shall report the delinquent one to the Board of Directors, by whom the person's name may be stricken from the rolls and all privileges of membership withdrawn Such person may, however, be re-instated by the Board of Directors upon payment of the arrears of dues.

ARTICLE III

Salaries

1 The Institute shall not pay a salary to any Officer, Director, or member of any committee

ARTICLE IV

Amendments

1 These By-Laws may be amended in the same manner as the Constitution or by a majority vote at any regularly convened meeting of the Institute, if the proposed amendment has been previously approved by the Board of Directors.

THE INSTITUTE OF MATHEMATICAL STATISTICS

Application for Junior Membership

Date _____

I am an undergraduate and hereby apply for Junior Membership in the Institute of Mathematical Statistics

Signature _____

Name _____
(Please print)

Mailing Address _____

College or University _____

Sponsored by _____

Applications and dues should be mailed to

EDWIN G. OLDS, *Treasurer*
Carnegie Institute of Technology
Pittsburgh, Pennsylvania

THE ANNALS
of
MATHEMATICAL STATISTICS

(FOUNDED BY H. C. CARVER)

THE OFFICIAL JOURNAL OF THE INSTITUTE OF MATHEMATICAL STATISTICS

Contents

PAGE

On Transformations Used in the Analysis of Variance. J. H. CURTISS 107

On Fundamental Systems of Probabilities of a Finite Number of Events. KAI LAI CHUNG 122

On the Efficient Design of Statistical Investigations. ABRAHAM WALD 134

Some Significance Tests for Normal Bivariate Distributions. D. S. VILLARS AND T. W. ANDERSON 141

Symmetric Tests of the Hypothesis that the Mean of One Normal Population Exceeds that of Another. HERBERT A. SIMON 149

On Indices of Dispersion. PAUL G. HOEL 155

On Serial Numbers. E. J. GUMBEL 163

Fitting General Gram-Charlier Series. PAUL A. SAMUELSON 179

A Method of Testing the Hypothesis that Two Samples are from the Same Population. HAROLD C. MATHISEN 188

Notes:

Note on the Independence of Certain Quadratic Forms. ALLEN T. CRAIG 195

A Characterization of the Normal Distribution. IRVING KAPLANSKY 197

News and Notices 199

Special Courses in Statistical Quality Control 202

Report on the New York Meeting of the Institute 204

Vol. XIV, No. 2 — June, 1943

THE ANNALS OF MATHEMATICAL STATISTICS

EDITED BY

S. S. WILKS, *Editor*

A. T. CRAIG — H. HOTELLING
W. E. DEMING — J. NEYMAN
T. C. FRY — W. A. SHEWHART

WITH THE COÖPERATION OF

W. G. COCHRAN — P. S. DWYER — P. G. HOEL
J. H. CURTISS — C. EISENHART — W. G. MADOW
H. F. DODGE — W. K. FELLER — A. WALD

The ANNALS OF MATHEMATICAL STATISTICS is published quarterly by the Institute of Mathematical Statistics, Mt. Royal & Guilford Aves., Baltimore, Md. Subscriptions, renewals, orders for back numbers and other business communications should be sent to the ANNALS OF MATHEMATICAL STATISTICS, Mt. Royal & Guilford Aves., Baltimore, Md., or to the Secretary of the Institute of Mathematical Statistics, E. G. Olds, Carnegie Institute of Technology, Pittsburgh, Pa. Changes in mailing address which are to become effective for a given issue should be reported to the Secretary on or before the 15th of the month preceding the month of that issue. The months of issue are March, June, September and December. Because of war-time difficulties of publication, issues may often be from two to four weeks late in appearing. *Subscribers are therefore requested to wait at least 30 days after month of issue before making inquiries concerning non-delivery.*

Manuscripts for publication in the ANNALS OF MATHEMATICAL STATISTICS should be sent to S. S. Wilks, Fine Hall, Princeton, New Jersey. Manuscripts should be typewritten double-spaced with wide margins, and the original copy should be submitted. Footnotes should be reduced to a minimum and whenever possible replaced by a bibliography at the end of the paper; formulae in footnotes should be avoided. Figures, charts, and diagrams should be drawn on plain white paper or tracing cloth in black India ink twice the size they are to be printed. Authors are requested to keep in mind typographical difficulties of complicated mathematical formulae.

Authors will ordinarily receive only galley proofs. Fifty reprints without covers will be furnished free. Additional reprints and covers furnished at cost.

The subscription price for the ANNALS is $5.00 per year. Single copies $1.50. Back numbers are available at $5.00 per volume, or $1.50 per single issue.

COMPOSED AND PRINTED AT THE
WAVERLY PRESS, INC.
BALTIMORE, MD., U. S. A.

Entered as second-class matter at the Post Office at Baltimore, Maryland, under the Act of March 3, 1879.

ON TRANSFORMATIONS USED IN THE ANALYSIS OF VARIANCE

By J. H. Curtiss
Cornell University

1. Introduction. Transformations of variates to render their distributions more tractable in various ways have long been used in statistics [12, chapter XVI] The present extensive use of the analysis of variance, particularly as applied to data derived from designs such as randomized blocks and Latin squares, has placed new emphasis on the usefulness of such transformations. In the more usual significance tests associated with the analysis of variance, it is assumed *a priori* that the plot yields are statistically independent normally distributed variates which all have the same variance, but which have possibly different means The hypotheses to be tested are then concerned with relations among these means. But in practice, it sometimes seems appropriate to specify for each variate a distribution in which the variance depends functionally upon the mean; moreover, in such cases, the specification is generally not normal. For example, when the data is in the form of a series of counts or percentages, a Poisson exponential or binomial specification may seem in order, and the variance of either of these distributions is functionally related to the mean of the distribution Before applying the usual normal theory to such data, it is clearly desirable to transform each variate so that normality and a stable variance are achieved as nearly as possible

Various transformations have been devised to do this, and a number of articles explaining the nature and use of these transformations have recently been published [1] However, the available literature on the subject appears to be mainly descriptive and non-mathematical. The object of this paper is to provide a general mathematical theory (sections 2 and 3) for certain types of transformations now in use In the framework of this theory we shall discuss in particular the square root and inverse sine transformations (section 4), and also several logarithmic transformations (section 4 and section 5).

2. General theory. As it arises in the analysis of variance, the problem of stabilizing a variance functionally related to a mean may be stated as follows: Suppose X is a variate whose mean $\mu = E(X)$ is a real variable with a range S of possible values, and whose standard deviation $\sigma = \sigma_X = \sigma(\mu)$ is a function of μ not identically constant Required, to find a function $T = f(X)$ such that both $f(X)$ and $\sigma_T^2 = E\{[T - E(T)]^2\}$ are functionally independent of μ for μ on S (By "functionally independent," we mean that $\frac{\partial f}{\partial \mu} \equiv 0$, and $\frac{\partial \sigma_T^2}{\partial \mu} \equiv 0$ for μ on S.)

[1] See references [1], [2], [3], [4], [5], [6], [13], [16]

The following line of argument is adopted in certain of the references mentioned above ([1], [2], [3], [4]): From the relation $dT = f'(X)dX$, we deduce as an approximation by some sort of summation process that $\sigma_T = f'(\mu)\sigma(\mu)$. Setting this expression equal to a constant, say c, we obtain $f'(\mu) = c/\sigma(\mu)$, so $f(x)$ is an indefinite integral of $c/\sigma(x)$. The roughness of the approximation used here is only too apparent.[2] For example, if X is normally distributed, then the variance of $T = X^2$ as given by the approximation is $4\sigma^2\mu^2$, while actually it is $4\sigma^2\mu^2 + 2\sigma^4$.

Indeed, it is easily seen that in important special cases the problem of stabilization as above stated could have no solution other than the trivial one in which T is identically constant on the set of points of increase of the d f.[3] of X. For instance, if X has a Poisson exponential distribution, then the identity $E[\{f(X) - E[f(X)]\}^2] \underset{\mu}{\equiv} c$, or $E\{[f(X)]^2\} \underset{\mu}{\equiv} c + \{E[f(X)]\}^2$, becomes

$$\sum_{k=0}^{\infty} [f(k)]^2 \frac{e^{-\mu}\mu^k}{k!} \underset{\mu}{\equiv} c + \left[\sum_{k=0}^{\infty} [f(k)] \frac{e^{-\mu}\mu^k}{k!}\right]^2, \qquad \mu > 0.$$

Expanding both sides in powers of μ, we need only equate the coefficients of the zero-th power of μ on each side to find that $[f(0)]^2 = c + [f(0)]^2$, which implies that $c = 0$ and hence that $f(0) = f(1) = f(2) = \cdots$ A similar demonstration can be given for the case in which X has a binomial distribution with a fixed number of values of the variate.

As to the problem of choosing $T = f(X)$ so that its distribution is exactly normal, we can observe immediately that a single-valued function $f(X)$ will never transform a variate X with a discrete distribution into a variate with a continuous one. On the other hand, any variate X with a continuous d.f. $F(x)$ can be transformed into a normally distributed variate T by the transformation $T = f(X)$ defined by the equation

$$F(X) = \int_{-\infty}^{T} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}\, dt.$$

However, aside from the practical difficulty of solving this equation for T, the resulting function $T = f(X)$ will not generally be functionally independent of the mean of X.

These considerations lead us to seek asymptotic solutions to the problems of normalization and stabilization. Such solutions are considered in the next section.

3. Asymptotic theorems. In the remainder of this paper, we shall suppose that the distribution of X depends on a parameter n which is to tend somehow to

[2] Tippett [14] says "This derivation is not mathematically sound, and the result is only justified if on application it is found to be satisfactory."

[3] i.e., distribution function. For any given one-dimensional variate X we shall denote the probability or relative frequency assigned to a set R by $P(R)$. The d.f of the variate then is the point function $F(x) = P(X \leq x)$. This function is sometimes called the cumulative frequency function of X.

infinity. The mean $\mu = \mu_n$ of X, with range S_n, will in general depend upon n (although by this we do not mean to exclude the case in which μ_n is constant for all values of n), and perhaps will depend also on some further independent parameters, which we shall denote collectively by θ, with range Σ. We shall seek a variate $T = f(X)$, in which $f(X)$ is functionally independent of μ and of the parameters θ for μ on S_n, θ on Σ, and such that the distribution of $f(X) - f(\mu_n)$ tends as $n \to \infty$ to a normal distribution, while $\lim_{n\to\infty}\sigma_T^2 = c^2$, where c^2 is an absolute constant. It is implied here that in case the additional parameters θ are present, the function $f(X)$ may depend non-trivially on n; but if n is the only parameter on which the distribution of X depends, then $f(X)$ must be functionally independent of n.

A solution to the problem just proposed is given in certain cases by Theorems 3.1 and 3.2 below, which are suggested by the heuristic reasoning of the second paragraph of section 2.

THEOREM 3.1. *Let $\psi_n(x)$ be a non-negative function of x and n, defined almost everywhere and integrable*[4] *with respect to x over any finite interval of the x-axis for each $n > 0$. Let*

$$T = f(X) = \int_a^X \psi_n(x)\,dx,$$

where a is an arbitrary constant. Let $F_n(y)$ be the d.f. of the variate $Y = (X - \mu_n)\psi_n(\mu_n)$. Suppose further that a continuous d.f. $F(y)$ exists such that $\lim_{n\to\infty}F_n(y) = F(y)$ for all values of y. Then either one of the following two conditions is a sufficient condition for the d.f. $H_n(w)$ of the variate $W = f(X) - f(\mu_n)$ to tend uniformly to $F(w)$, $-\infty < w < \infty$:

(a) *To each w for which $0 < F(w) < 1$, there corresponds for all n sufficiently large at least one root $x = x_n$ to the equation*

$$\int_{\mu_n}^{x} \psi_n(u)\,du = w, \tag{3.1}$$

and this root x_n has the property that

$$\lim_{n\to\infty}(x_n - \mu_n)\psi_n(\mu_n) = w. \tag{3.2}$$

(b) *For all n sufficiently large, $\psi_n(\mu_n) > 0$, and $\lim_{n\to\infty}q_n(w) = 1$ uniformly in any closed finite subinterval of the open interval defined by $0 < F(w) < 1$, where*

$$q_n(w) = \frac{\psi_n(w[\psi_n(\mu_n)]^{-1} + \mu_n)}{\psi_n(\mu_n)}. \tag{3.3}$$

To prove this theorem we shall first suppose that condition (a) is satisfied. Let w_1 and w_2 be the end points of the open interval (possibly infinite) defined by $0 < F(w) < 1$. If w lies in this interval, and if n is large enough for the root x_n in (3.1) to exist, then from the monotonic character of $\int_{\mu_n}^{X}\psi_n(x)\,dx$ we can

[4] "Integrable" here means absolutely integrable in the sense of Lebesgue.

infer that

$$H_n(w) = P[f(X) - f(\mu_n) \leqq w] = P\left[\int_{\mu_n}^{x} \psi_n(x)\,dx \leqq w\right]$$

$$= P(X \leqq x_n) = P[Y \leqq (x_n - \mu_n)\psi_n(\mu_n)] \tag{3.4}$$

$$= F_n[(x_n - \mu_n)\psi_n(\mu_n)].$$

Since $F(w)$ is continuous, $\lim_{n\to\infty} F_n(w) = F(w)$ *uniformly* on any finite or infinite interval of values of w, as is well known.[5] Therefore $\lim_{n\to\infty} F_n(w_n) = F(w)$ if $\lim_{n\to\infty} w_n = w$. Thus from (3.2) and (3.4), we find that $\lim_{n\to\infty} H_n(w) = F(w)$ for $w_1 < w < w_2$.

If $w' \leqq w_1$, and $w_1 < w'' < w_2$, then $0 \leqq H_n(w') \leqq H_n(w'') = F(w'') + [H_n(w'') - F(w'')]$. We can make the right hand member of this relation less than any given positive number ϵ by first choosing w'' so that $F(w'') < \frac{1}{2}\epsilon$ (it will be remembered that $F(w)$ is a continuous d.f., and $F(w_1) = 0$) and then choosing n so large that the quantity in square brackets is also less than $\frac{1}{2}\epsilon$ in absolute value. Thus $\lim_{n\to\infty} H_n(w') = 0$ Similarly if $w' \geqq w_2$, we can show that $\lim_{n\to\infty} H_n(w') = 1$ Hence $\lim_{n\to\infty} H_n(w) = F(w)$ for all w, and it follows that the limit is uniform on any finite or infinite interval of values of w.

We shall now show that condition (a) in the theorem is a consequence of condition (b). The result follows at once from the following simple lemma:

LEMMA. *If $\gamma_n(w)$ is a non-negative function integrable over any finite interval of values of w; and if $\lim_{n\to\infty}\gamma_n(w) = 1$ uniformly in any finite closed subinterval of an interval $w_1 < w < w_2$, then for every value of w in this interval there exists for all n sufficiently large a solution $y = y_n$ of the equation $\int_0^y \gamma_n(z)\,dz = w$, and the solution y_n has the property that $\lim_{n\to\infty} y_n = w$.*

For it is clear that if w satisfies the inequality $w_1 < w < w_2$, and if $\eta > 0$ be chosen so that $w_1 < w - \eta < w + \eta < w_2$, then for all n sufficiently large,

$$\int_0^{w-\eta} \gamma_n(z)\,dz \leqq w \leqq \int_0^{w+\eta} \gamma_n(z)\,dz.$$

Thus for each n sufficiently large, there exists a root y_n of the equation $\int_0^y \gamma_n(z)\,dz = w$, and furthermore, this root satisfies the inequality $w - \eta \leqq y_n \leqq w + \eta$. Since η is arbitrarily small, the proof of the lemma is complete.

To apply the lemma, we make the change of variables $z = (u - \mu_n)\psi_n(\mu_n)$ in the integral in (3.1), which reduces it to the form

$$\int_0^y q_n(z)\,dz, \qquad y = (x - \mu_n)\psi_n(\mu_n), \tag{3.5}$$

and the conclusion that (a) is implied by (b) now follows at once

[5] See [7], Theorem 11, pp. 29–30, also [8].

We add the remark that the uniformity of the limit of $q_n(z)$ in condition (b) may be replaced by the condition that for each closed finite sub-interval there exists a function $q(w)$ which dominates $q_n(w)$ for all n sufficiently large.

Our second theorem, which is stated in the terminology and notation of Theorem 3.1, is concerned with the limit of the variance of $T = f(X)$. From the mere fact that the distribution of W tends to a limiting form, it by no means follows that the mean and variance of the distribution of W approach those of the limiting form, as may be shown by trivial examples Thus additional hypotheses on $\psi_n(x)$ and on the behavior of the distribution of Y become necessary.

THEOREM 3 2. *Let T (or $f(X)$), Y, $F_n(y)$ and $F(y)$ be defined as in Theorem 3 1 Let the mean and variance of the distribution defined by $F(y)$ exist and have respective values 0 and c^2 Then the following three conditions, taken together, are sufficient that*

$$\lim_{n\to\infty}[E(T) - f(\mu_n)] = 0, \tag{3.6}$$

$$\lim_{n\to\infty}\sigma_T^2 = c^2: \tag{3 7}$$

(i) *$E(Y^2)$ exists for $n > 0$, and $\lim_{n\to\infty}E(Y^2) = c^2$.*

(ii) *Condition (b) of Theorem 3 1 holds*

(iii) *$f(Y[\psi_n(\mu_n)]^{-1} + \mu_n) - f(\mu_n) = O \mid Y \mid$ uniformly in n as $\mid Y \mid \to \infty$*

As a preliminary step in the proof, we observe that (i) and the relations $\lim_{n\to\infty}F_n(y) = F(y)$, $c^2 = \int_{-\infty}^{+\infty} y^2\, dF(y)$, imply that the improper integral $\int_{-\infty}^{+\infty} y^2\, dF_n(y)$ converges uniformly in n for $n > 0$ As the integrand is positive, the following result is equivalent to the uniform convergence of the integral: *For every $\epsilon > 0$, there exist numbers A_1 and A_2, $A_1 < A_2$, such that for all n sufficiently large,*

$$\left(\int_{-\infty}^{A_1} + \int_{A_2}^{\infty}\right) y^2\, dF_n(y) < \epsilon.$$

To prove this, we write

$$\left(\int_{-\infty}^{A_1} + \int_{A_2}^{\infty}\right) y^2\, dF_n(y) = [E(Y^2) - c^2] + \left[\int_{A_1}^{A_2} y^2\, dF(y) - \int_{A_1}^{A_2} y^2\, dF_n(y)\right] + \left[c^2 - \int_{A_1}^{A_2} y^2\, dF(y)\right].$$

We first choose A_1 and A_2 so that the last bracket here is less than $\frac{1}{2}\epsilon$ in absolute value By condition (i), the first bracket approaches zero as n tends to infinity, and the Helly-Bray theorem [10, p 15] states that the second bracket also approaches zero as n tends to infinity, so for all n sufficiently large, the sum of the first two brackets is in absolute value less than $\frac{1}{2}\epsilon$.

It is important to notice that we can always choose A_1 and A_2 in the above

demonstration so that $A_1 > w_1$, $A_2 < w_2$, where w_1 and w_2 are as usual the endpoints of the interval defined by $0 < F(w) < 1$.

To continue with the proof of the theorem, we remark that by a change of variables similar to the one used to derive (3.5), the function $W = f(X) - f(\mu_n)$ may be expressed as a function of Y in the following manner:

$$W = \int_{\mu_n}^{X} \psi_n(x)\, dx = \int_0^Y q_n(w)\, dw = Q_n(Y),$$

where $q_n(w)$ is given by (3.3). In terms of W, (3.6) and (3.7) become, respectively,

$$\lim_{n\to\infty} E(W) = 0, \tag{3.8}$$

$$\lim_{n\to\infty} \{E(W^2) - [E(W)]^2\} = c^2, \tag{3.9}$$

and these are the equations which we now establish.

Conditions (ii) and (iii) obviously imply that $\lim_{n\to\infty} Q_n(y) = y$ uniformly in any finite closed subinterval of the interval $w_1 < y < w_2$, and that a constant M exists such that $|Q_n(y)| \leqq M|y|$ for all n. If $E(Y^2)$ exists, so will $E(Y)$ Now

$$\begin{aligned} E(W) &= \int_{-\infty}^{+\infty} Q_n(y)\, dF_n(y) \\ &= \int_{-\infty}^{+\infty} Q_n(y)\, dF_n(y) - \int_{-\infty}^{+\infty} y\, dF_n(y) \\ &= \left(\int_{-\infty}^{A_1} + \int_{A_2}^{\infty}\right)[Q_n(y) - y]\, dF_n(y) + \int_{A_1}^{A_2} [Q_n(y) - y]\, dF_n(y), \end{aligned}$$

where $w_1 < A_1 < A_2 < w_2$. Therefore

$$|E(W)| \leqq \left(\int_{-\infty}^{A_1} + \int_{A_2}^{\infty}\right)(M+1)|y|\, dF_n(y) + \int_{A_1}^{A_2} |Q_n(y) - y|\, dF_n(y).$$

From the uniform convergence of $\int_{-\infty}^{+\infty} y^2 dF_n(y)$, proved above, we can conclude that the pair of improper integrals in this inequality can be made less than an arbitrary $\frac{1}{2}\epsilon > 0$ by proper choice of A_1 and A_2. The third integral approaches zero, by the general Helly-Bray Theorem [10, p. 16], and so becomes less than $\frac{1}{2}\epsilon$ for all n sufficiently large. Thus we have established (3.8) To show that (3.9) is true, we have merely to prove that $\lim_{n\to\infty} E(W^2) = c^2$. Since $E(Y^2) = \int_{-\infty}^{+\infty} y^2 dF_n(y)$, we may write

$$E(W^2) - c^2 = \int_{-\infty}^{+\infty} \{[Q_n(y)]^2 - y^2\}\, dF_n(y) + [E(Y^2) - c^2].$$

The integral may be shown to approach zero by the argument used in the case of $E(W)$, and the required result then follows from condition (1) of the theorem. The proof is now complete.

The sufficient conditions in Theorem 3.2 can be modified in various more or less obvious ways The existence of the limiting d.f. $F(y)$ was essentially used in the proof only to secure the uniform convergence of $\int_{-\infty}^{+\infty} y^2 dF_n(y)$. Condition (ii) can again be modified along the lines suggested at the end of the proof of Theorem 3.1 Condition (iii) was used only to secure the uniform convergence of the integral $\int_{-\infty}^{+\infty} [Q_n(y)]^2 dF_n(y)$.

For later reference, we shall supplement Theorems 3 1 and 3.2 with the following simple result, which is practically self-evident.

THEOREM 3 3. *Let the distribution of a variate Y depend upon a parameter n, let $F_n(y)$ be the d.f. of Y, and let $F(y)$ be a continuous d.f with the property that* $\lim_{n\to\infty} F_n(y) = F(y)$. *Let a_n be a function of n such that* $\lim_{n\to\infty} a_n = a \neq 0$. *Then the d.f. of the variate $Z = a_n Y$ tends as $n \to \infty$ to the d.f. $F(z/a)$ if $a > 0$, and to the d f. $1 - F(z/a)$ if $a < 0$. If the variance of Y exists and tends to c^2 as $n \to \infty$, then the variance of $a_n Y$ tends to a^2c^2 as $n \to \infty$.*

If $F(y)$ is the d.f. of a reduced normal distribution, i e.,

$$F(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-\frac{1}{2}t^2}\, dt,$$

then $F(z/a)$ is also the d.f. of a normal distribution with mean zero and variance a^2. More generally, any affine transformation of a normal variate yields another normal variate.

4. Applications. The theorems of the preceding section have the effect of referring the properties of the distribution of the transformation $T = f(X)$ of Theorem 3 1 back to those of the distribution of a related variate Y In the applications given in the present section, we shall let $\psi_n(\mu_n)$ be proportional to the reciprocal of the standard deviation of X. The theorems of section 3 state in this case that if the reduced, or standardized, distribution of X approaches a limiting form, then under certain circumstances, the distribution of $f(X) - f(\mu_n)$ will approach a similar limiting form, and σ_T^2 will approach a quantity independent at least of n. In the applications considered here, the reduced distribution of X will always approach the reduced normal distribution.

(I) **The square root transformation for a variate with a Poisson exponential distribution.** *Let X have a Poisson exponential distribution with parameter n. If α is an arbitrary constant, and if*

$$T = f(X) = \begin{cases} \sqrt{X + \alpha}, & X \geqq -\alpha \\ 0, & X < -\alpha \end{cases} \tag{4.1}$$

then the distribution of $T - \sqrt{n+\alpha}$ *tends as* $n \to \infty$ *to a normal distribution which has mean zero and variance* $\frac{1}{4}$, *and* $\lim_{n\to\infty}\sigma_T^2 = \frac{1}{4}$. For $\mu_n = n$, $\sigma_X = \sqrt{n}$, and it is well known[6] that the distribution of the reduced variate $(X - n)/\sqrt{n}$ tends to the reduced normal distribution as $n \to \infty$. By Theorem 3.3, the distribution of the variate

$$Y = \frac{X-n}{2\sqrt{n+\alpha}} = \frac{1}{2}\cdot\sqrt{\frac{n}{n+\alpha}}\cdot\frac{X-n}{\sqrt{n}},$$

will tend to normality as $n \to \infty$, and the variance of Y will tend to the value $\frac{1}{4}$, which is also the variance of the limiting distribution. Setting

$$\psi_n(x) = \begin{cases} \dfrac{1}{2\sqrt{x+\alpha}}, & x > -\alpha \\ 0, & x \leqq -\alpha, \end{cases}$$

we obtain from $T = f(X) = \int_{-\alpha}^{X} \psi_n(x)dx$ the formula given in (4.1). To prove the statement in italics, we must show that conditions (ii) and (iii) of Theorem 3.2 are satisfied. We have, assuming $n > -\alpha$,

$$q_n(w) = \begin{cases} \left(1 + \dfrac{2w}{\sqrt{n+\alpha}}\right)^{-1}, & w > -\frac{1}{2}\sqrt{n+\alpha} \\ 0, & w \leqq -\frac{1}{2}\sqrt{n+\alpha}, \end{cases}$$

so clearly (ii) is satisfied. Also,

$$W = f(Y[\psi_n(\mu_n)]^{-1} + \mu_n) - f(\mu_n)$$
$$= \begin{cases} \sqrt{2Y\sqrt{n+\alpha} + n + \alpha} - \sqrt{n+\alpha}, & Y > -\frac{1}{2}\sqrt{n+\alpha} \\ -\sqrt{n+\alpha}, & Y \leqq -\frac{1}{2}\sqrt{n+\alpha}, \end{cases}$$

from which it follows at once that $|W| < 2|Y|$ for all Y, and so (iii) is satisfied.

The degree of approximation involved in the equation $\lim_{n\to\infty}\sigma_T^2 = \frac{1}{4}$ has been investigated numerically by Bartlett [1] for values of n from .5 to 15.0 in the cases $\alpha = 0$ and $\alpha = \frac{1}{2}$. He found that the variance of $\sqrt{X + (\frac{1}{2})}$ is considerably closer to the limit $(\frac{1}{4})$ for $1 \leqq n \leqq 10$ than is the variance of $\sqrt{X}$. At $n = 15$, the variance of $\sqrt{X}$ is .256, and that of $\sqrt{X + (\frac{1}{2})}$ is .248.

The question of the degree of convergence to normality and of the possibility of selecting an optimum value of α remain open. By expanding the function $\sqrt{X+\alpha}$ in a Taylor series about $X = n$ with remainder in the form due to Schlömilch, it is possible to derive as accurate an estimate of $|\sigma_T^2 - (\frac{1}{4})|$ as may

[6] See (e.g.) [9].

be desired. A rough result easily obtainable by this method is that $|\sigma_T^2 - (\frac{1}{4})| \leqq 3/(4n)$, $n > 0$.

(II) **The square root transformation for a variate with a Γ distribution.** *Let X have a distribution whose density function is of the following type:*

$$\varphi(x) = \begin{cases} 0 & x \leqq 0 \\ Kx^{\frac{1}{2}n-1}e^{-hx}, & x \geqq 0,\ h > 0. \end{cases} \tag{4.2}$$

If α is an arbitrary constant, and if

$$T = f(X) = \begin{cases} \sqrt{X + \alpha}, & X \geqq -\alpha \\ 0\ , & X < -\alpha, \end{cases} \tag{4.3}$$

then the distribution of $T - \sqrt{(n/2h) + \alpha}$ tends as $n \to \infty$ to a normal distribution which has mean zero and variance $1/4h$, and $\lim_{n\to\infty}\sigma_T^2 = 1/(4h)$ For $\mu_n = n/(2h)$, $\sigma_X = \sqrt{n}/(h\sqrt{2}) = \sqrt{\mu_n/h}$ The distribution of the reduced variate tends to normality as $n \to \infty$,[7] so that of the variate

$$Y = \frac{x - \mu_n}{2\sqrt{\mu_n + \alpha}} = \frac{1}{2}\sqrt{\frac{n}{nh + 2h^2\alpha}} \cdot \frac{x - \mu_n}{\sqrt{\mu_n/h}}$$

tends to normality also with limiting variance $1/(4h)$. Setting

$$\psi_n(x) = \begin{cases} \dfrac{1}{2\sqrt{x + \alpha}}, & x > -\alpha \\ 0\ , & x \leqq -\alpha, \end{cases}$$

we obtain T in (4.3) from the relation $T = \int_{-\alpha}^{x} \psi_n(x)\,dx$. The work of verifying that the conditions of Theorem 3.2 are satisfied is the same as in the case of the Poisson exponential distribution treated above, and will not be repeated.

For example, if s^2 denotes the variance of a random sample of $n + 1$ observations drawn from a normal parent distribution with variance σ^2, then it is well known that $(n + 1)s^2$ is distributed according to (4.2) with $h = 1/(2\sigma^2)$. We thus can deduce the further facts, also well known, that the distribution of $\sqrt{n + 1}\ s - \sigma\sqrt{n}$ tends to normality, and that the variance of $s\sqrt{n + 1}$ approaches the limiting value $\frac{1}{2}\sigma^2$. If n is an integer and $h = \frac{1}{2}$, the distribution defined by (4.2) is called a χ^2 distribution with n degrees of freedom, and the variate is often denoted by χ^2. Our conclusion in this case is that the distribution of $\sqrt{2\chi^2} - \sqrt{2n}$ tends to a normal one with zero mean and unit variance. From this result and the fact that $\sqrt{2\ n-\ 1} - \sqrt{2n} = O(n^{-\frac{1}{2}})$, it follows immediately that $\sqrt{2\chi^2} - \sqrt{2n - 1}$ has the same limiting distribution as $\sqrt{2\chi^2} - \sqrt{2n}$. This result,[8] due to Fisher, is familiar to all users of his table of the probability levels of χ^2.

[7] See (e g.) [9].

[8] For a discussion of the degree of convergence involved here, see [9].

(III) **The inverse sine transformation for a binomial variate.** *Let X have a binomial relative frequency distribution with parameter p and the n values $0, 1/n, 2/n, \cdots, n/n$. If α is an arbitrary constant, and if*

$$T = f(X) = \begin{cases} \sqrt{n}\sin^{-1}\sqrt{X + \frac{\alpha}{n}}, & -\frac{\alpha}{n} \leqq X \leqq 1 - \frac{\alpha}{n} \\ 0, & X < -\frac{\alpha}{n}, \quad X > 1 - \frac{\alpha}{n}, \end{cases} \tag{4.4}$$

where T is measured in radians, then the distribution of $T - \sqrt{n}\sin^{-1}\sqrt{p + (\alpha/n)}$ tends as $n \to \infty$ to a normal distribution which has mean zero and variance $\frac{1}{4}$, and $\lim_{n\to\infty}\sigma_T^2 = \frac{1}{4}$. For here, $\mu_n = p$, and $\sigma_X^2 = pq/n$, where $q = 1 - p$; and the familiar DeMoivre-Laplace theorem states that the distribution of the reduced variate $\sqrt{n}(X - p)/\sqrt{pq}$ will tend to normality as $n \to \infty$. Hence by Theorem 3.3 the distribution of

$$Y = \frac{\sqrt{n}(X - p)}{2\sqrt{\left(p + \frac{\alpha}{n}\right)\left(q - \frac{\alpha}{n}\right)}} \tag{4.5}$$

will tend to normality with a limiting variance of $\frac{1}{4}$, which is also the variance of the limiting distribution. Setting

$$\psi_n(x) = \begin{cases} \dfrac{\sqrt{n}}{2\sqrt{\left(x + \frac{\alpha}{n}\right)\left(1 - x - \frac{\alpha}{n}\right)}}, & -\frac{\alpha}{n} < x < 1 - \frac{\alpha}{n} \\ 0 & x \leqq -\frac{\alpha}{n}, x \geqq 1 - \frac{\alpha}{n}, \end{cases}$$

we obtain (4.4) from the integral

$$T = \int_{-\alpha/n}^{X} \psi_n(x)\,dx.$$

In proving the conditions (ii) and (iii) of Theorem 3.2 are satisfied, we shall assume for simplicity that $\alpha = 0$. We find that

$$q_n(w) = \begin{cases} \left(1 + 2w\dfrac{q - p}{\sqrt{npq}} - \dfrac{4w^2}{n}\right)^{-\frac{1}{2}}, & -\frac{1}{2}\sqrt{\frac{np}{q}} < w < \frac{1}{2}\sqrt{\frac{nq}{p}} \\ 0, & w \leqq -\frac{1}{2}\sqrt{\frac{np}{q}}, \quad w \geqq \frac{1}{2}\sqrt{\frac{nq}{p}}, \end{cases}$$

so obviously (ii) is satisfied. From the Law of the Mean in the form due to Schlömilch, we have

$$\begin{aligned} W &= \sqrt{n}\sin^{-1}\sqrt{p + 2\sqrt{\frac{pq}{n}}\,Y} - \sqrt{n}\sin^{-1}\sqrt{p} \\ &= 2Y\left[\frac{1 - \theta}{\left(1 + 2\theta\sqrt{\frac{q}{np}}\,Y\right)\left(1 - 2\theta\sqrt{\frac{p}{nq}}\,Y\right)}\right]^{\frac{1}{2}}, \end{aligned} \tag{4.6}$$

$$0 < \theta < 1, \qquad -\frac{1}{2}\sqrt{\frac{np}{q}} < Y < \frac{1}{2}\sqrt{\frac{nq}{p}}.$$

The denominator of the coefficient of $2Y$ here is a quadratic function of Y with a negative coefficient of Y^2, and so must assume its least value in the Y range indicated in (4.6) at one end or the other of the range From this it is readily seen that the coefficient of $2Y$ is actually always less than unity For values of Y outside the range, the second member of (4.6) indicates that $W = O(\sqrt{n}) = O(Y)$. Hence (iii) is satisfied, and the proof of the statement in italics is complete for the case $\alpha = 0$. The more general case presents no important new difficulties

In practice, it is often convenient to express X as a percentage. This merely has the effect of multiplying Y in (4 5) by 100 We find in this case that $\sqrt{n}\sin^{-1}\sqrt{X + 100\alpha/n} - \sqrt{n}\sin^{-1}\sqrt{100p + 100\alpha/n}$ has a distribution approaching normality, and $\sigma_T \to 50$ instead of $\frac{1}{2}$.

Bartlett [1] gives numerical results in the cases $n = 10$, $\alpha = 0$ and $n = 10$, $\alpha = \frac{1}{2}$, which indicate that perhaps the choice $\alpha = \frac{1}{2}$ is more suitable if the estimated p is near 0 or 1, but the choice $\alpha = 0$ is preferable if the estimated p lies between 3 and .7 However, there seems to be no good reason to believe that these conclusions should be valid for other values of n. The question of an optimum α, and of the degree of convergence to normality remain open. We note in passing that the latter problem could doubtless be profitably studied by combining the methods of proof of Theorem 3 1 with the results of Uspensky [15, pp 129–130] on the degree of approximation of the reduced binomial d.f. to the normal d.f.

IV. **Other transformations of a binomial variate.** *Let X have a binomial relative frequency distribution with the parameter p and the n values $0, 1/n, 2/n, \cdots, n/n$.*

(a) *If*

$$T = f(X) = \begin{cases} \sqrt{n}\sinh^{-1}\sqrt{X} = \sqrt{n}\log\left(\sqrt{X} + \sqrt{1+X}\right),^{9} & X \geqq 0 \\ 0, & X < 0, \end{cases}$$

then the distribution of $T - \sqrt{n}\sinh^{-1}\sqrt{p}$ tends as $n \to \infty$ to a normal distribution which has mean zero and variance $q/(4+4p)$, and $\lim_{n\to\infty}\sigma_T^2 = q/(4+4p)$.

(b) *If*

$$T = f(X) = \begin{cases} \sqrt{n}\log X, & X > 0, \\ 0, & X \leqq 0, \end{cases}$$

then the distribution of $T - \sqrt{n}\log p$ tends as $n \to \infty$ to a normal distribution which has mean zero and variance q/p, and $\lim_{n\to\infty}\sigma_T^2 = q/p$.

(c) *If*

$$T = f(X) = \begin{cases} \frac{1}{2}\sqrt{n}\log\dfrac{X}{1-X}, & 0 < X < 1, \\ 0, & X \leqq 0, \quad X \geqq 1, \end{cases}$$

[9] All logarithms in this paper are to the base e.

then the distribution of $T - \frac{1}{2}\sqrt{n}\log\frac{p}{1-p}$ *tends as* $n \to \infty$ *to a normal distribution which has mean zero and variance* $1/(4pq)$, *and* $\lim_{n\to\infty}\sigma_T^2 = 1/(4pq)$

Since the limiting variance of each of these transformations involves the parameter p, they are not to be regarded as solutions of the problem of asymptotic variance stabilization proposed at the beginning of section 3, although it is perhaps of some interest that their distributions become asymptotically normal.

In case (a), $f'(x) = \sqrt{n}/(2\sqrt{x^2+x})$, $x > 0$. Setting $\psi_n(x) = f'(x)$, $x > 0$, and $\psi_n(x) = 0$, $x \leqq 0$, we obtain

$$Y = (X-p)\psi_n(p) = \frac{\sqrt{n}(X-p)}{\sqrt{pq}}\cdot\frac{\sqrt{q}}{2\sqrt{1+p}}, \tag{4.7}$$

and this variate obviously has the limiting distribution ascribed to $T - \sqrt{n}\sinh^{-1}\sqrt{p}$ in the statement in italics. The truth of that statement now follows by an argument similar to that used in the case of the inverse sine transformation.

If p is allowed to vary with n in such a way that $\lim_{n\to\infty} np = \infty$, it is known that the reduced distribution of X will still tend to normality.[10] If we suppose that $\lim_{n\to\infty} p = 0$, but $\lim_{n\to\infty} np = \infty$, we find from Theorem 3.3 that the limiting distribution of Y in (4.7) will be normal with mean zero and variance $\frac{1}{4}$, and that $\sigma_Y^2 \to \frac{1}{4}$. It is easily verified that the conditions (ii) and (iii) of Theorem 3.2 are still satisfied, so we find that the limiting distribution of $[\sqrt{n}\sinh^{-1}\sqrt{X} - \sqrt{n}\sinh^{-1}\sqrt{p}]$ is normal, with mean zero and variance $\frac{1}{4}$, and $\sigma_T^2 \to \frac{1}{4}$. However, since n is now the only independent parameter, we cannot here regard the transformation $T = \sqrt{n}\sinh^{-1}\sqrt{X}$ as a solution of the problem of variance stabilization, because the variate T depends explicitly upon n

If in case (b) we proceed as in case (a), we obtain as the analogue of (4.7) the formula

$$Y = (X-p)\psi_n(p) = \frac{\sqrt{n}(X-p)}{\sqrt{pq}}\sqrt{\frac{q}{p}},$$

and this variate has the limiting distribution ascribed to $T - \sqrt{n}\log X$ in the statement in italics. It now turns out that although condition (ii) of Theorem 3.2 is satisfied, condition (iii) is not satisfied. We are then faced with the problem of proving directly that the improper integral

$$\int_{-\sqrt{n}}^{+\infty} [\sqrt{n}\log(p + py/\sqrt{n}) - \sqrt{n}\log p]^2\, dF_n(y)$$

converges uniformly.[11] The trouble occurs only at the lower limit of integration, and may be resolved by first integrating by parts, then dividing the range

[10] See (e.g.) [9]

[11] See the remarks following the proof of Theorem 3.2.

$(-\sqrt{n}, A_1)$ into two ranges $(-\sqrt{n}, -\log n)$ and $(-\log n, A_1)$, and then applying Uspensky's results [15, pp. 129–130], on the degree of approximation involved in the DeMoivre-Laplace theorem.

Case (c) may be handled in a similar manner.

5. The logarithmic transformation. We shall suppose throughout this section that X is a variate whose mean μ_n and standard deviation σ in the relation $\sigma = k_n(\mu_n + \alpha)$, where α is an arbitrary constant, $k_n > 0$, and $\lim_{n\to\infty} k_n$ exists and is finite If k_n is constant for all n, say $k_n = k > 0$, and if we use the heuristic argument of the second paragraph of section 2 to attempt to find a transformation which will stabilize the variance of X at k^2, we arrive at the function $T = \log (X + \alpha)$, $X > -\alpha$ It is the purpose of this section to study the asymptotic properties of this transformation.

The theory of such a transformation differs in certain important respects from that of the transformations considered in sections 3 and 4 For one thing, our starting point in the study of each transformation considered in section 4 was the fact that although $P(X < 0) = 0$, nevertheless the reduced distribution of X tended to normality as $n \to \infty$. But in the present case, if X is a variate such that $P(X \leqq -\alpha) = 0$, then the corresponding reduced variate $Y = (X - \mu_n)/[k_n(\mu_n + \alpha)]$ has a d.f. $F_n(y)$ such that $F_n(-1/k_n) = 0$. Thus if $\lim_{n\to\infty} k_n = k > 0$, the limiting distribution of Y, if it exists, must have a d.f. $F(y)$ such that $F(-1/k - 0) = 0$. Therefore the limiting distribution of Y can never be normal if $k > 0$.

Moreover (in contrast to the situation in Theorem 3.1) if the reduced variate Y does have a limiting distribution, the variate

$$(5.1)\quad W = \frac{1}{k_n}\log (X + \alpha) - \frac{1}{k_n}\log (\mu_n + \alpha) = \int_{\mu_n}^{X} \frac{1}{k_n(u + \alpha)}\,du, \quad X > -\alpha$$

may have a limiting distribution which is not the same as that of Y. More specifically, we have the following result:

THEOREM 5 1 *Let* $P(X \leqq -\alpha) = 0$, *let* $\lim_{n\to\infty} k_n = k \geqq 0$, *let* $F_n(y)$ *be the d f of the reduced variate*

$$Y = \frac{X - \mu_n}{k_n(\mu_n + \alpha)},$$

and let $H_n(w)$ *be the d.f. of the variate* W *given by* (5.1). *If a continuous d f.* $F(y)$ *exists such that* $\lim_{n\to\infty} F_n(y) = F(y)$ *for all* y, *then*

$$\lim_{n\to\infty} H_n(w) = \begin{cases} F\left[\dfrac{e^{kw} - 1}{k}\right], & k > 0 \\ F(w), & k = 0. \end{cases}$$

The proof is simpler than the statement, essentially we have only to notice that

$$H_n(w) = P\left[-\frac{1}{k_n} < Y \leqq \frac{e^{k_n w} - 1}{k_n}\right] = F_n\left[\frac{e^{k_n w} - 1}{k_n}\right], \qquad -\infty < w < \infty,$$

and apply the reasoning used above in connection with (3.4).

From the study of the distribution of T, we now turn for a moment to the question of the limit if σ_T^2. Here the situation is more consistent with the results of section 3.

THEOREM 5.2. *Under the hypotheses of Theorem* 5.1 *and under the additional conditions that the improper integral* $\int_{-\infty}^{0} w^2\, dH_n(w)$ $\left(\text{or } \int_{-1/k_n}^{0} k_n^{-2}[\log(1 + k_n y)]^2\, dF_n(y)\right)$ *converges uniformly in* n *and that* $\int_{-\infty}^{+\infty} y^2\, dF(y) = 1 = E(Y^2)$, *the following relations hold:*

$$\lim_{n\to\infty} E(W) = \begin{cases} \int_{-1/k}^{\infty} \frac{1}{k} \log(1 + ky)\, dF(y), & k > 0, \\ 0, & k = 0, \end{cases} \tag{5.2}$$

$$\lim_{n\to\infty} E(W^2) = \begin{cases} \int_{-1/k}^{\infty} \frac{1}{k^2} [\log(1 + ky)]^2\, dF(y), & k > 0 \\ 1, & k = 0 \end{cases} \tag{5.4}$$

The variance σ_T^2 of the variate $T = \log(X + \alpha)$ is related to these mean values by the equation $\sigma_T^2 = k_n^2\{E(W^2) - [E(W)]^2\}$. Thus if $F(y)$ is independent of any unknown parameters θ, and if k is positive and is presumed to have the same value for all variates in any given problem, then the transformation $T = \log(X + \alpha)$ is seen to yield an asymptotic stabilization of the variance under the conditions of Theorem 5.2. If $k = 0$, we find from either Theorem 5.2 or the proof of Theorem 5.2 that $T = \log(X + \alpha)$ converges stochastically to $\log(\mu_n + \alpha)$.

The proof of Theorem 5.2 is similar to that of Theorem 3.2 and will be omitted.

Theorem 5.1 raises the following question: Just what limiting distribution must Y have if $k > 0$, in order that the distribution of W tend to normality? To answer this, we shall note the following simple non-asymptotic result:

THEOREM 5.3. *A necessary and sufficient condition that* X *have a continuous distribution with density function*

$$\varphi(x) = \begin{cases} \frac{1}{\sqrt{2\pi \log(k^2 + 1)}} \frac{1}{x + \alpha} \times \exp\left[\frac{-\left(\log \frac{(x + \alpha)\sqrt{k^2 + 1}}{u + \alpha}\right)^2}{2 \log(k^2 + 1)}\right], & x > -\alpha \\ 0, & x \leqq -\alpha \end{cases} \tag{5.4}'$$

for which $\sigma_X = k(\mu + \alpha)$, *is that the variate* $T = \log(X + \alpha)$ *have a normal distribution with mean* $\log(\mu + \alpha) - \log\sqrt{k^2 + 1}$ *and variance* $\log(k^2 + 1)$.

The proof may be given by a routine change of variables.[12] It is to be noticed that the heuristic argument of the second paragraph of section 2 would lead to the incorrect result that the variance of T was k^2 instead of $\log(k^2 + 1)$. In case $k = 1$, the mean and variance of T are respectively $\log(\mu + \alpha) - .347$ and .693. If the transformation $T = \log_{10}(X + \alpha)$ is used, the new mean is $\log_{10}(\mu + \alpha) - \log_{10}\sqrt{k^2 + 1}$ and the new variance is $.189 \log(k^2 + 1$, which for values of k near zero has the approximate value $.189k^2$.[13]

If X is distributed according to (5.4), the density function $F'(y)$ of the corresponding reduced variate $Y = (X - \mu)/[k(\mu + \alpha)]$ is

$$(5.5)\quad F'(y) = \begin{cases} \dfrac{k}{\sqrt{2\pi \log(k^2 + 1)}} \cdot \dfrac{1}{1 + ky} \\ \qquad \times \exp\left[-\dfrac{\{\log[(1 + ky)\sqrt{k^2 + 1}]\}^2}{2 \log(k^2 + 1)} \right], & y > -\dfrac{1}{k} \\ 0 & y \leqq -\dfrac{1}{k}. \end{cases}$$

The d f of the variate $W = k^{-1}[\log(X + \alpha) - \log(\mu + \alpha)]$ is $F[(e^{kw} - 1)/k]$, and, of course, the distribution of W is normal with mean $-k^{-1}\log\sqrt{k^2 + 1}$, and variance $k^{-2}\log(k^2 + 1)$. These are the respective values of the integrals in (5.2) and (5.3).

If now the distribution of X depends on a parameter n in such a way that as $n \to \infty$, the distribution of the corresponding reduced variate $Y = (X - \mu_n)/[k_n(\mu_n + \alpha)]$ tends to the distribution given by (5.5), it follows from the above remarks and from Theorem 5.1 that the variate W given by (5.1) has a normal limiting distribution. Furthermore, under the uniform convergence condition of Theorem 5.2, it follows that σ_T^2 tends to the value $\log(k^2 + 1)$, where $T = \log(X + \alpha)$.

These facts provide a sound mathematical basis for the use of the logarithmic transformation, which has had a long history of empirical success in problems of normalization [12, chapter XVI] and stabilization ([6], [16]). When it appears from a reasonably large number of observations on a variate (which is essentially bounded from below) that the standard deviation of the variate is proportional to the mean, then a possible specification for the variate is a distribution of the form (5.4); or, at least for large values of μ, it may be assumed that the distribution of the reduced variate is given by (5.5). Then the variate $T = \log(X + \alpha)$, where $-\alpha$ is any number less than the lower bound of X, will be exactly or approximately normally distributed with a variance independent of the value of μ.

Since (5.4) is only one of an infinity of various different types of distribution

---

[12] Finney [11] has considered the problem of efficiently estimating the variance of the X of Theorem 5.3 in the case $\alpha = 0$. (The actual density function (5.4) appears nowhere in his paper.)

[13] Given (without explanation) by Cochran [6, p. 165].

in which the mean and standard deviation are proportional, the user of a logarithmic transformation in the analysis of variance should always apply tests for departure from normality to the observed distribution of T values. From the point of view of specification, the situation here would seem to be less reassuring than in the cases considered in section 4 While it is true that the Poisson exponential distribution is only one of many types of distribution in which the variance and mean are equal, nevertheless the specification of a Poisson distribution can generally be preceded by a fairly strong chain of *a priori* inductive reasoning This would not seem to be the case in the specification of (5.4). Theorems 5.1 and 5.2 furnish some grounds for a suspicion that the logarithmic transformation may possibly be more successful in stabilizing the variance than in normalizing the data. The burden of proof, however, lies with the experimenter.[14]

REFERENCES

[1] M. S Bartlett, "The square root transformation in the analysis of variance," *Jour Roy. Stat Soc. Suppl.*, Vol. 3 (1936), pp. 68–78

[2] Geoffrey Beall, "The transformation of data from entomological field experiments so that the analysis of variance becomes applicable," *Biometrika*, Vol. 32 (1942), pp 243–262

[3] C. I. Bliss, "The transformation of percentages for use in the analysis of variance," *Ohio Jour Science*, Vol 38 (1938), pp. 9–12.

[4] A. Clark and W H. Leonard, "The analysis of variance with special reference to data expressed as percentages," *Jour. Amer. Soc. Agron.*, Vol. 31 (1939), pp 55–56

[5] W. G. Cochran, "The analysis of variance when experimental errors follow the Poisson or binomial laws," *Annals of Math Stat.*, Vol. 9 (1940), pp 335–347.

[6] W. G. Cochran, "Some difficulties in the statistical analysis of replicated experiments," *Empire Jour. Expl Agric.*, Vol. 6 (1938), pp. 157–175.

[7] H. Cramér, *Random Variables and Probability Distributions*, Cambridge, 1937

[8] J H Curtiss, "A note on the theory of moment generating functions," *Annals of Math Stat*, Vol 13 (1942), pp. 430–433.

[9] J. H Curtiss, "Convergent sequences of probability distributions," *Am. Math Monthly*, Vol. 50 (1943), pp 94–105

[10] G. C. Evans, *The Logarithmic Potential*, New York, 1927.

[11] D J. Finney, "On the distribution of a variate whose logarithm is normally distributed," *Jour. Roy. Stat. Soc Suppl*, Vol. 7 (1941), pp 155–161

[12] Arne Fisher, *The Mathematical Theory of Probabilities*, Second edition, New York, 1930

[13] R. A. Fisher and F Yates, *Statistical Tables*, London, 1938.

[14] L H C. Tippett, "Statistical methods in textile research. Part 2, Uses of the binomial and Poisson distributions," *Shirley Inst. Mem*, Vol. 13 (1934), pp. 35–72.

[15] J. V. Uspensky, *Introduction to Mathematical Probability*, New York, 1937.

[16] C. B. Williams, "The use of logarithms in the interpretation of certain entymological problems," *Annals of Appl. Biol*, Vol. 24 (1937), pp. 404–414

[14] A transformation closely related to the logarithmic one is $T = k^{-1} \sinh^{-1}(kX)^{\frac{1}{2}}$, where k is an estimate of the Charlier coefficient of disturbancy of a Poisson distribution This transformation has recently been studied from an empirical point of view by Beall [2], it was suggested by the heuristic argument of section 2 applied to the case in which $\sigma^2 = \mu + k\mu^2$. Beall presents evidence that for the particular data which he considered, the transformation seemed to stabilize the variance and normalize A mathematical theory would follow the lines laid down above in the case of $T = \log(X + \alpha)$.

ON FUNDAMENTAL SYSTEMS OF PROBABILITIES OF A FINITE NUMBER OF EVENTS

By Kai Lai Chung

Tsing Hua University, Kunming, China

We consider a probability function $P(E)$ defined over the Borel set of events generated by the n arbitrary events $E_1, \cdots, E_n$, which will be denoted by $\mathfrak{L}(1, \cdots, n)$.

We use the same notations as in the author's former paper[1], with the following abbreviations. We denote a combination $(\alpha_1 \cdots \alpha_a)$ simply by (α), and use the corresponding Latin letter a for its number of members. Similarly we write (β) for $(\beta_1 \cdots \beta_b)$, but (ν) for $(1, \cdots, n)$ We say that (β) belongs to (α) and write $(\beta) \in (\alpha)$ when and only when the set $(\beta_1 \cdots \beta_b)$ is a subset of $(\alpha_1 \cdots \alpha_a)$. Then and then only we write $(\alpha) - (\beta)$ for the subset of elements of (α) that do not belong to (β); thus we may write it as (γ) with $c = a - b$. When and only when (α) and (β) have no common elements, we write $(\alpha) + (\beta)$ for the set of elements that belong either to (α) or to (β); thus we may write it as (γ), with $c = a + b \leqq n$. We note the case for empty sets: $(0) + (0) = (0)$. Now we can write $p_{[(\alpha)]}$ for $p_{[\alpha_1 \cdots \alpha_a]}$, $p_{((\alpha))}$ for $p_{\alpha_1 \cdots \alpha_a}$, $p_b((\alpha))$ for $p_b(\alpha_1 \cdots \alpha_a)$, etc. Further we denote by $p_{[b]}((\alpha))$ $(1 \leqq b \leqq a \leqq n)$ the probability of the occurrence of exactly b events out of $E_{\alpha_1}, \cdots, E_{\alpha_a}$, and write

$$P_a^{(m)}((\nu)) = \sum_{(\alpha) \in (\nu)} p_m((\alpha)), \qquad P_a^{[m]}((\nu)) = \sum_{(\alpha) \in (\nu)} p_{[m]}((\alpha));$$

since a is fixed by the left-hand sides, the summations on the right-hand sides are to be extended to all the $\binom{n}{a}$-combinations of (ν).

A sum written $\sum_{(\beta) \in (\alpha)}$ is to be extended to all combinations (β), $b = 0, 1, \cdots, a$ belonging to (α), when b is not previously fixed; it is to be extended to all the $\binom{a}{b}$-combinations belonging to (α), when b is previously fixed.

Definition 1. *A system of quantities is said to form a fundamental system of probabilities for a set of events if and only if the probability of every event in the set can be expressed in terms of these quantities.*

Definition 2. *An event in $\mathfrak{L}(1, \cdots, n)$ is said to be symmetrical if and only if it is identical with every event obtained by interchanging any pair of suffixes (i, j) $(i, j = 1, \cdots, n)$ in the definition of it.* The subset of symmetrical events in $\mathfrak{L}(1, \cdots, n)$ will be denoted by $\mathfrak{S}(1, \cdots, n)$.

From the normal form[2] of every event in $\mathfrak{L}(1, \cdots, n)$ and the principle of

---

[1] "On the probability of the occurrence of at least m events among n arbitrary events," *Annals of Math. Stat.*, Vol. 12, 1941.

[2] See Hilbert-Ackermann, *Grundzüge der theoretischen Logik*, Chap. 1.

total probabilities, we can easily see the truth of the following theorems, which may of course be made more precise.

THEOREM. *The system of* $p_{[(\alpha)]}$, $(\alpha) \in (\nu)$, 2^n *in number, forms a fundamental system for* $\mathfrak{L}(1, \cdots, n)$.

THEOREM. *The system of* $p_{[a]}((\nu))$, $0 \leqq a \leqq n$, $n+1$ *in number, forms a fundamental system for* $\mathfrak{S}(1, \cdots, n)$.

Next, a theorem of Broderick[3], in a less precise form, may be stated:

The system of $p_{((\alpha))}$ $(p_{((0))} = 1)$, $(\alpha) \in (\nu)$, 2^n *in number, forms a fundamental system for* $\mathfrak{L}$.

We may add in an easy way the following

THEOREM. *The system of* $S_a((\nu))$ $S_0((\nu)) = 1$, $0 \leqq a \leqq n$, $n+1$ *in number, forms a fundamental system for* $\mathfrak{S}$.

In the present paper we shall prove, *inter alia*, the following four theorems of the above type, stated in more precise forms

THEOREM 1. *For any* E *in* $\mathfrak{L}$, *we have*

$$P(E) = c_0 + \sum_{\substack{(\alpha) \in (\nu) \\ \alpha \neq 0}} c_\alpha p_1((\alpha)),$$

where $c_0 = 0$ *or* 1 *and the* c_α*'s are integers; and they are unique*[4].

THEOREM 2. *For any* E *in* $\mathfrak{S}$, *we have*

$$P(E) = c_0 + \sum_{a=1}^{n} c_a P_a^{(1)},$$

where $c_0 = 0$ *or* 1 *and the* c_a*'s are integers; and they are unique.*

THEOREM 3. *For any* E *in* $\mathfrak{L}$, *we have*

$$P(E) = d_0 + \sum_{\substack{(\alpha) \in (\nu) \\ \alpha \neq 0}} d_\alpha p_{[1]}((\alpha)),$$

where $d_0 = 0$ *or* 1 *and the* d_α*'s are rational numbers and they are unique.*

THEOREM 4. *For any* E *in* $\mathfrak{S}$, *we have*

$$P(E) = d_0 + \sum_{a=1}^{n} d_a p_a^{[1]},$$

where $d_0 = 0$ *or* 1 *and the* d_a*'s are rational numbers; and they are unique.*

Less precisely, we may say that the system of $p_1((\alpha))$ or $p_{[1]}((\alpha))$ forms a fundamental system for $\mathfrak{L}$; the system of $P_a^{(1)}((\nu))$ or $P_a^{[1]}((\alpha))$ forms a fundamental system for $\mathfrak{S}$.

In fact however, we shall give much more than the mere proofs of

[3] Fréchet, "Compléments à un théorème de T. S. Broderick concernant les événements dependants," *Proc. Edinburgh Math. Soc.*, Ser. 2, Vol. 6 (1939).

[4] "Unique" in the sense that it is impossible to replace therein the coefficients c by other numbers which are independent of the Borel set of events and the probability function.

these theorems. We shall establish the following explicit formulas for the general parameter m.

$$\text{(i)} \qquad p_{[(0)]} = 1 - p_1((\nu)),$$

$$\text{(1.1)} \quad \text{(ii)}' \qquad p_{[(\alpha)]} = \sum_{\substack{(\beta)\,\epsilon\,(\alpha) \\ n-a+b\neq 0}} (-1)^{b-1} p_1((\nu) - (\alpha) + (\beta)),^{5} \qquad 1 \leq a \leq n.$$

$$\text{(1)} \qquad p_{[(\alpha)]} = (-1)^m \frac{m-1}{n-1} \sum_{c=m}^{n} \sum_{d=\max(0,c-a)}^{\min(c,n-a)} (-1)^{c-d} \binom{n-2}{a+d-m}^{-1} \sum_{\substack{(\delta)\,\epsilon\,(\nu)-(\alpha) \\ (\gamma)-(\delta)\,\epsilon\,(\alpha)}} p_m((\gamma) - (\delta) + (\delta)), \qquad n \geq a \geq m \geq 2.^{5}$$

$$\text{(2.1)} \qquad p_{[a]}((\nu)) = \sum_{\substack{b=n-a \\ b\neq 0}}^{n} (-1)^{b-n+a} \binom{b}{n-a} P_b^{(1)}((\nu)), \qquad 1 \leq a \leq n.$$

$$\text{(2)} \qquad p_{[a]}((\nu)) = \sum_{b=m}^{n} (-1)^{b-m} L(n, a, b, m) P_b^{(m)}((\nu)), \qquad n \geq a \geq m \geq 2,$$

where

$$L(n, a, b, m) = \begin{cases} 0, & b < n - a + m - 1, \\ (-1)^{n-a} \binom{a}{m-1}^{-1}, & b = n - a + m - 1, \\ \dfrac{(-1)^{n-a}(m-1)!\,(b-m)!\,(a-m)!\,\{ab - n(m-1)\}}{a!\,(n-a)!\,(a+b-n-m+1)!}, & b > n - a + m - 1. \end{cases}$$

$$\text{(3)} \quad \text{(i)} \qquad p_{[(0)]}((\nu)) = 1 - \frac{1}{n} \sum_{c=1}^{n} \binom{n-1}{c-1}^{-1} P_c^{[1]}.$$

$$\text{(ii)} \qquad p_{[(\alpha)]} = (-1)^m \frac{m}{n} \sum_{c=m}^{n} \sum_{d=\max(0,c-a)}^{\min(c,n-a)} (-1)^{c-d} \binom{n-1}{a+d-m}^{-1} \sum_{\substack{(\delta)\,\epsilon\,(\nu)-(\alpha) \\ (\gamma)-(\delta)\,\epsilon\,(\alpha)}} p_{[m]}((\gamma) - (\delta) + (\delta)), \qquad n \geq a \geq m \geq 1.$$

$$\text{(4)} \qquad p_{[a]}((\nu)) = \sum_{b=m+n-a}^{n} (-1)^{n-a+b-m} \binom{b-m}{n-a} \binom{a}{m}^{-1} P_b^{[m]}((\nu)), \qquad n \geq a \geq m \geq 1.$$

A simpler derivation of (1) than that given in an earlier paper[1] follows. Let us write Poincaré's formula as follows:

$$p_m((\beta)) = \sum_{c=m}^{b} (-1)^{c-m} \binom{c-1}{m-1} S_c((\beta)).$$

---

[5] Obviously we mean $((\nu) - (\alpha)) + (\beta)$ and $((\gamma) - (\delta)) + (\delta)$ respectively; similarly in the sequel.

Then for a fixed $b \geqq m$, summing over all $(\beta) \in (\nu)$, we get

$$\sum_{(\beta)\in(\nu)} p_m((\beta)) = \sum_{c=m}^{b} (-1)^{c-m} \binom{c-1}{m-1}\binom{n-c}{b-c} S_c((\nu)).$$

Hence

$$\begin{aligned}\sum_{b=m}^{n} (-1)^{b-m} \sum_{(\beta)\in(\nu)} p_m((\beta)) &= \sum_{c=m}^{n} \binom{c-1}{m-1} S_c((\nu)) \sum_{b=c}^{n} (-1)^{b-c} \binom{n-c}{b-c} \\ &= \sum_{c=m}^{n} \binom{c-1}{m-1} S_c((\nu)) \begin{cases} 1 & \text{if } c = n \\ 0 & \text{if } c < n \end{cases} \\ &= \binom{n-1}{m-1} S_n((\nu)) = \binom{n-1}{m-1} p((\nu)).\end{aligned} \tag{1}$$

A change of notation gives, for $a + b \geqq m$,

$$\binom{a+b-1}{m-1} p_{((\alpha)+(\beta))} = \sum_{c=m}^{a+b} (-1)^{c-m} \sum_{(\gamma)\in(\alpha)+(\beta)} p_m((\gamma)).$$

Hence

$$\begin{aligned}&\binom{a+b-1}{m-1} \sum_{(\beta)\in(\nu)-(\alpha)} p_{((\alpha)+(\beta))} \\ &\quad = \sum_{c=m}^{a+b} (-1)^{c-m} \sum_{d=\max(0,c-a)}^{\min(c,n-a)} \binom{n-a-d}{b-d} \sum_{\substack{(\delta)\in(\nu)-(\alpha) \\ (\gamma)-(\delta)\in(\alpha)}} p_m((\gamma) - (\delta) + (\delta)).\end{aligned}$$

Substituting in the well-known formula, for $a \geqq 1$

$$p_{[(\alpha)]} = \sum_{b=0}^{n-a} (-1)^b \sum_{(\beta)\in(\nu)-(\alpha)} p_{((\alpha)+(\beta))},$$

we get for $n \geqq a \geqq m$

$$\begin{aligned}p_{[(\alpha)]} = \sum_{c=m}^{n} (-1)^{c-m} &\sum_{d=\max(0,c-a)}^{\min(c,n-a)} \\ &\sum_{\substack{(\delta)\in(\nu)-(\alpha) \\ (\gamma)-(\delta)\in(\alpha)}} p_m((\gamma) - (\delta) + (\delta)) \left\{ \sum_{b=0}^{n-a} (-1)^b \binom{n-a-d}{b-d} \binom{a+b-1}{m-1}^{-1} \right\}.\end{aligned} \tag{1}$$

Thus the problem reduces to the summation of the following series:

$$\sum_{b=0}^{n-a} (-1)^b \binom{n-a-d}{b-d} \binom{a+b-1}{m-1}^{-1}.$$

Case 1: $m = 1$. In this case the series reduces to

$$\sum_{b=0}^{n-a} (-1)^b \binom{n-a-d}{b-d} = \begin{cases} (-1)^{n-a} & \text{if } d = n-a, \\ 0 & \text{if } d < n-a. \end{cases}$$

Hence for $a \geqq 1$,

$$p_{[(\alpha)]} = \sum_{c=\max(1,\,n-a)}^{n} (-1)^{c-1} \sum_{(\gamma)-((\nu)-(\alpha))\,\epsilon\,(\alpha)} p_1((\nu)-(\alpha)+(\gamma)-((\nu)-(\alpha)))(-1)^{n-a}$$

Writing $(\gamma) - ((\nu) - (\alpha)) = (\beta)$, we obtain

$$p_{[(\alpha)]} = \sum_{b=\max(1-n+a,\,0)}^{a} (-1)^{b-1} \sum_{(\beta)\,\epsilon\,(\alpha)} p_1((\nu) - (\alpha) + (\beta)).$$

This is equivalent to (1.1), (ii), while (i) is trivial.

Case 2: $m \geqq 2$. We have, for $c \geqq 1$,

$$\sum_{l=0}^{a} (-1)^l \binom{a}{l} \binom{b+l}{c}^{-1} = \frac{c}{a+b} \binom{a+b-1}{b-c}^{-1},$$

which is easily proved by induction on a.

Hence for $m \geqq 2$,

$$\begin{aligned}
\sum_{b=0}^{n-a} (-1)^b \binom{n-a-d}{b-d} \binom{a+b-1}{m-1}^{-1} &= \sum_{b'=-d}^{n-a-d} (-1)^{d+b'} \binom{n+a-d}{b'} \binom{a+b'+d-1}{m-1}^{-1} \\
&= (-1)^d \sum_{b'=0}^{n-a-d} (-1)^b \binom{n-a-d}{b'} \binom{a+d-1+b'}{m-1}^{-1} \\
&= (-1)^d \frac{m-1}{n-1} \binom{n+2}{a+d-m}^{-1}
\end{aligned}$$

Substituting in (1) we get formula (1).

To derive formula (2.1) for a fixed a, $1 \leqq a \leqq n$, we sum (1.1, ii), which gives

$$p_{[a]}((\nu)) = \sum_{(\alpha)\,\epsilon\,(\nu)} p_{[(\alpha)]} = \sum_{\substack{b=0 \\ n-a+b\neq 0}}^{a} (-1)^{b-1} \sum_{(\alpha)\,\epsilon\,(\nu)} \sum_{(\beta)\,\epsilon\,(\alpha)} p_1((\nu) - (\alpha) + (\beta)).$$

Letting $(\nu) - (\alpha) + (\beta) = (\gamma)$, we get

$$p_{[a]}((\nu)) = \sum_{c=\max(1,\,n-a)}^{n} (-1)^{n-a+c-1} \binom{c}{n-a} \sum_{(\gamma)\,\epsilon\,(\nu)} p_1((\gamma)),$$

which is formula (2.1).

The following form of Poincaré's formula is of assistance in deriving (2):

$$p_{[a]}((\nu)) = \sum_{c=a}^{n} (-1)^{c-a} \binom{c}{a} S_a((\nu)).$$

Substituting from (1), we get

$$p_{[a]}((\nu)) = \sum_{c=a}^{n} (-1)^{c-a}\binom{c}{a}\binom{c-1}{m-1}^{-1} \sum_{b=m}^{c} (-1)^{b-m}\binom{n-b}{c-b} P_b^{(m)}((\nu))$$

$$= \sum_{b=m}^{n} (-1)^{b-m} P_b^{(m)}((\nu)) \left\{ \sum_{c=\max(a,b)}^{n} (-1)^{c-a}\binom{c}{a}\binom{n-b}{c-b}\binom{c-1}{m-1}^{-1} \right\}.$$

Thus the problem reduces to the summation of the following series:

$$L(n, a, b, m) = \sum_{c=\max(a,b)}^{n} (-1)^{c-a}\binom{c}{a}\binom{n-b}{c-b}\binom{c-1}{m-1}^{-1}$$

First, we have, for $z \geqq 0$, $y \geqq w$,

$$\sum_{x=\max(0,1-w)}^{z} (-1)^x\binom{z}{x}(x+y)\ \cdots\ (x+w)$$

$$= \begin{cases} 0 & \text{if } y - w + 1 < z, \\ \dfrac{(-1)^z y!(y+1-w)!}{(z+w-1)!(y+1-w-z)!} & \text{if } y - w + 1 \geqq z, \end{cases}$$

which may be easily proved by induction on z
Next, we have

$$L(n, a, b, m) = \frac{(m-1)!}{a!} \sum_{c=\max(a,b)}^{n} (-1)^{c-a}\binom{n-b}{c-b}\frac{c(c-m)!}{(c-a)!}$$

$$= \frac{(m-1)!}{a!} \sum_{c'=\max(0,a-b)}^{n-b} (-1)^{c'+b-a}\binom{n-b}{c'}\frac{(c'+b)(c'+b-m)!}{(c'+b-a)!}$$

$$= (-1)^{b-a}\frac{(m-1)!}{a!} \sum_{c'=\max(0,a-b)}^{n-b} (-1)^{c'} \cdot \binom{n-b}{c'}\frac{(c'+b-m+1)! + (m-1)(c'+b-m)!}{(c'+b-a)!}$$

$$= (-1)^{b-a}\frac{(m-1)!}{a!}\{T(n, a, b, m) + (m-1)T(n, a, b, m+1)\},$$

where

$$T(n, a, b, m) = \sum_{c=\max(0,a-b)}^{n-b} (-1)^c\binom{n-b}{c}\frac{(c+b-m+1)!}{(c+b-a)!}$$

$$= \begin{cases} 0 & \text{if } b < n - a + m - 1, \\ \dfrac{(-1)^{n-b}(a-m+1)!(b-m+1)!}{(n-a)!(a+b-n-m+1)!} & \text{if } b \geqq n - a + m - 1, \end{cases}$$

by the preceding formula. Thus we get the explicit expression for $L(n, a, b, m)$ given in formula (2), which is thereby proved

The derivations of formulas 3 and 4 are similar to the above and may be omitted.

Now we can give the essential argument for Theorems 1–4. It is evident that for any E in $\mathfrak{L}$, we have

$$P(E) = \Sigma p_{[(\alpha)]},$$

where the summation extends to certain combinations $(\alpha) \in (\nu)$. Substituting from formula (1.1) we get Theorem 1; substituting from formula (3) we get Theorem 3. Next, for any E in $\mathfrak{S}$, we have

$$P(E) = \Sigma p_{[a]}((\nu)),$$

where the summation extends to certain values of a. Substituting from formula (1 1), (i) and formula (2) we get Theorem 2; substituting from formula (3), (i) and formula (4) we get Theorem 4. We may note these proofs are "constructive".

It remains to prove the uniqueness of the coefficients in Theorems 1–4. For Broderick's theorem this has been done by Fréchet[3], by introducing "independent events" Our proof will be based on the conditions of existence, also initiated by Fréchet[6], for the systems $p_1((\alpha))$, $p_{[1]}((\alpha))$, $P_a^{(1)}((\nu))$, $P_a^{[1]}((\nu))$.

The conditions of existence of the system $p_1((\alpha))$ have been given by the author in the paper[1], though the proof there is not quite complete.

1 *Conditions of existence of the system* $P_a^{(1)}((\nu))$ Given n quantities $Q_a^{(1)}$, $1 \leqq a \leqq n$; what are the necessary and sufficient conditions that they may be the system of $P_a^{(1)}((\nu))$'s, $1 \leqq a \leqq n$, of a probability function defined over $\mathfrak{S}(1, \cdots, n)$?

From formula (1.1), (i) and formula (2) it is evident that necessary conditions are, for $1 \leqq a \leqq n$,

$$\sum_{\substack{b=n-a \\ b \neq 0}}^{n} (-1)^{b-n+a-1} \binom{b}{n-a} Q_b^{(1)} \geqq 0, \tag{3}$$

$$1 - Q_n^{(1)} \geqq 0,$$

and

$$\sum_{a=1}^{n} \sum_{\substack{b=n-a \\ b \neq 0}}^{n} (-1)^{b-n+a-1} \binom{b}{n-a} Q_b^{(1)} + 1 - Q_n^{(1)} = 1. \tag{4}$$

The last condition can be re-written as

$$\sum_{b=1}^{n} (-1)^{b-1} Q_b^{(1)} \sum_{a=\max(1, n-b)}^{n} (-1)^{n-a} \binom{b}{n-a} + 1 - Q_n^{(1)} = 1,$$

which reduces to the identity $1 = 1$.

---

[6] "Conditions d'existence de système d'événements associés à certaines probabilités," *Jour. de Math.*, 1940. However, our interpretation of the term would mean instead "conditions of existence of a probability function defined over a Borel set of events, etc."

To show that the conditions (3) are sufficient, put

$$p_{[a]} = \sum_{b=n-a}^{n} (-1)^{b-n+a-1} \binom{b}{n-a} Q_b^{(1)},$$

$$p_{[0]} = 1 - Q_n^{(1)}.$$

By (3) and (4) we have, for $0 \leqq a \leqq n$,

$$p_{[a]} \geqq 0 \qquad \text{and} \qquad \sum_{a=0}^{n} p_{[a]} = 1.$$

Hence they are actually the $p_{[a]}((\nu))$'s of a probability function. We want to show that the $P_a^{(1)}((\nu))$'s of this probability function coincide with the given $Q_a^{(1)}$'s, so that this is the probability function we seek. We have,

$$P_b^{(1)}((\nu)) = \sum_{(\beta) \in (\nu)} p_1((\beta)) = \sum_{a=1}^{n} p_{[a]} \sum_{h=\max(1,b-n+a)}^{\min(a,b)} \binom{a}{h}\binom{n-a}{b-h}$$

$$= \sum_{c=0}^{n} \left\{ \sum_{a=\max(1,n-c)}^{n} (-1)^{c-n+a-1} \binom{c}{n-a} \sum_{h=\max(1,b-n+a)}^{\min(a,b)} \binom{a}{h}\binom{n-a}{b-h} \right\} Q_c^{(1)}.$$

Now the series in curl brackets

$$= \sum_{a=\max(1,n-c)}^{n-b} (-1)^{c-n+a-1} \binom{c}{a} \left\{ \binom{n}{b} - \binom{n-a}{b} \right\}$$

$$+ \sum_{a=n-b+1}^{n} (-1)^{c-n+a-1} \binom{c}{n-a} \binom{n}{b}$$

$$= \sum_{a=\max(1,n-c)}^{n} (-1)^{c-n+a-1} \binom{c}{n-a} \binom{n}{b}$$

$$- \sum_{a=\max(1,n-c)}^{n-b} (-1)^{c-n+a-1} \binom{c}{n-a} \binom{n-a}{b}.$$

If $c = n$, the last

$$= \binom{n}{b} - \sum_{a=1}^{n-b} (-1)^{a-1} \binom{n}{n-b} \binom{n-b}{a}$$

$$= \binom{n}{b} - \binom{n}{b} \sum_{a=1}^{n-b} (-1)^{a-1} \binom{n-b}{a} = \begin{cases} 1 & \text{if } b = n; \\ 0 & \text{if } b \neq n. \end{cases}$$

If $c < n$, we have

$$= 0 + (-1)^c \sum_{a=n-c}^{n-b} (-1)^{n-a} \binom{c}{n-a} \binom{n-a}{b}$$

$$= (-1)^c \sum_{a'=c}^{b} (-1)^{a'} \binom{c}{a'} \binom{a'}{b} = \begin{cases} 1 & \text{if } b = c; \\ 0 & \text{if } b \neq c. \end{cases}$$

Therefore

$$P_b^{(1)}((\nu)) = Q_b^{(\nu)}.$$

2. *Conditions of existence of the system* $p_{[1]}((\alpha))$. Given $2^n - 1$ quantities $q_{[1]}((\alpha))$, $(\alpha) \in (\nu)$, $a \geqq 1$, what are the necessary and sufficient conditions that they may be the system of $p_{[1]}((\alpha))$'s, of a probability function defined over $\mathfrak{L}(1, \cdots, n)$?

From formula 3 it is evident that necessary conditions are

$$\frac{1}{n}\sum_{c=1}^{n}\sum_{d=\max(0,c-a)}^{\min(c,n-a)}(-1)^{c-d-1}\binom{n-1}{a+d-1}^{-1}\sum_{\substack{(\delta)\in(\nu)-(\alpha)\\(\gamma)-(\delta)\in(\alpha)}} q_{[1]}((\gamma)-(\delta)+(\delta)) \geqq 0, \tag{5}$$

$$1-\frac{1}{n}\sum_{c=1}^{n}\binom{n-1}{c-1}^{-1}\sum_{(\gamma)\in(\nu)} p_{[1]}((\gamma)) \geqq 0;$$

and

$$1+\frac{1}{n}\sum_{(\alpha)\in(\nu)}\sum_{c=1}^{n}\sum_{d=\max(0,c-a)}^{\min(c,n-a)}(-1)^{c-d-1}\binom{n-1}{a+d-1}^{-1}\sum_{\substack{(\delta)\in(\nu)-(\alpha)\\(\gamma)-(\delta)\in(\alpha)}} q_{[1]}((\gamma)-(\delta)+(\delta)) = 1. \tag{6}$$

Consider the sum

$$\sum_{(\alpha)\in(\nu)}\sum_{d=\max(0,c-a)}^{\min(c,n-a)}(-1)^{d}\binom{n-1}{a+d-1}^{-1}\sum_{\substack{(\delta)\in(\nu)-(\alpha)\\(\gamma)-(\delta)\in(\alpha)}} q_{[1]}((\gamma)-(\delta)+(\delta)).$$

For a fixed (δ), the number of ways of writing $(\gamma) = (\gamma) - (\delta) + (\delta)$ is $\binom{c}{d}$, then since $(\gamma) - (\delta) \in (\alpha)$ but $(\alpha) - ((\gamma) - (\delta)) \in (\nu) - (\gamma)$, the number of choices of (α) is $\binom{n-c}{a-c+d}$. Thus the coefficient of $q_{[1]}((\gamma))$ in the sum is

$$\sum_{a=0}^{n}\sum_{d=\max(0,c-a)}^{\min(c,n-a)}(-1)^{d}\binom{c}{d}\binom{n-c}{a-c+d}\binom{n-1}{a+d-1}^{-1} = \binom{n-1}{c-1}^{-1}\sum_{a=0}^{n}\sum_{d=\max(0,c-a)}^{\min(c,n-a)}(-1)^{d}\binom{c}{d}\binom{a+d-1}{c-1} = 0.$$

Therefore the condition (6) reduces to the identity $1 = 1$.

To show that conditions (6) are sufficient, put the left-hand sides of (5) equal to $p_{[(\alpha)]}$ and $p_{[(0)]}$ respectively. Then

$$\begin{aligned} p_{((\alpha))} &= \sum_{(\beta)\in(\nu)-(\alpha)} p_{[(\alpha)+(\beta)]} \\ &= \frac{1}{n}\sum_{c=1}^{n}(-1)^{c-1}\sum_{b=0}^{n-a}\sum_{d=\max(0,c-a-b)}^{\min(c,n-a-b)}(-1)^{d}\binom{n-1}{a+d-1}^{-1}\sum_{(\beta)\in(\nu)-(\alpha)}\sum_{\substack{(\delta)\in(\nu)-(\alpha)-(\beta)\\(\gamma)-(\delta)\in(\alpha)+(\beta)}} q_{[1]}((\gamma)-(\delta)+(\delta)). \end{aligned} \tag{7}$$

Let $(\gamma) = (\gamma) - (\phi) + (\phi)$, where $(\phi) \epsilon (\alpha)$, $(\gamma) - (\phi) \epsilon (\nu) - (\alpha)$. Then the sum in the curl brackets can be written, by a combinatorial calculation, as

$$\sum_{f=0}^{\min(a,c)} \left\{ \sum_{b=0}^{n-a} \sum_{d=\max(0,c-f-b)}^{\min(c-f,n-a-b)} (-1)^d \binom{c-f}{d} \binom{n-a-c+f}{b-c+d+f} \binom{n-1}{a+b+d-1}^{-1} \right\} \sum_{\substack{(\phi)\epsilon(\alpha) \\ (\gamma)-(\phi)\epsilon(\nu)-(\alpha)}} q_{[1]}((\gamma) - (\phi) + (\phi)).$$

The sum in the last curl brackets is

$$\binom{n-1}{a+c-f-1}^{-1} \sum_{b=0}^{n-a} \sum_{d=\max(0,c-f-b)}^{\min(c-f,n-a-b)} (-1)^d \binom{c-f}{d} \binom{a+b+d-1}{a+c-f-1}.$$

Inverting the order of summations,

$$\binom{n-1}{a+c-f-1}^{-1} \sum_{d=\max(0,c-f-n+a)}^{\min(c-f,n-a)} (-1)^d \binom{c-f}{d} \sum_{b=c-f-d}^{n-a-d} \binom{a+b+d-1}{a+c-f-1}$$

$$= \binom{n-1}{a+c-f-1}^{-1} \sum_{d=\max(0,c-f-n+a)}^{\min(c-f,n-a)} (-1)^d \binom{c-f}{d} \binom{n}{a+c-f}$$

$$= \binom{n}{a+c-f} \binom{n-1}{a+c-f-1}^{-1} \sum_{d=0}^{c-f} (-1)^d \binom{c-f}{d} = \begin{cases} \dfrac{n}{a} & \text{if } f = c, \\ 0 & \text{if } f \neq c. \end{cases}$$

Hence (7) reduces to

$$p_{((\alpha))} = \frac{1}{a} \sum_{c=1}^{n} (-1)^{c-1} \sum_{(\gamma)\epsilon(\alpha)} q_{[1]}((\gamma)).$$

Then

$$S_b((\alpha)) = \sum_{(\beta),\epsilon(\alpha)} p_{((\beta))} = \frac{1}{b} \sum_{\substack{(\delta)\epsilon(\alpha) \\ d\neq 0}} (-1)^{d-1} \binom{a-d}{b-d} q_{[1]}((\delta))$$

$$p_{[1]}((\alpha)) = \sum_{b=1}^{a} (-1)^{b-1} S_b((\alpha))$$

$$= \sum_{\substack{(\delta)\epsilon(\alpha) \\ d\neq 0}} \left\{ \sum_{b=1}^{a} (-1)^{b-d} \binom{a-d}{b-d} \right\} q_{[1]}((\delta)) = q_{[1]}((\alpha)).$$

The conditions of existence of the system $P_a^{[1]}((\nu))$, $1 \leqq a \leqq n$, are similarly deduced from formula (3), (i) and formula (4) with $m = 1$.

Now we can prove the uniqueness of the coefficients in Theorems 1–4. Since the proofs are all exactly similar, we take Theorem 2. Suppose, if possible, there exists another system of coefficients c_a', $0 \leqq a \leqq n$ so that

$$P(E) = c_0 + \sum_{a=1}^{n} c_a P_a^{(1)}((\nu)) = c_0' + \sum_{a=1}^{n} c_a' P_a^{(1)}((\nu)).$$

Taking the difference, we get a linear polynomial in the variables $P_a^{(1)}((\nu))$, $1 \leqq a \leqq n$ which must vanish:

$$(8) \qquad (c_0 - c_0') + \sum_{a=1}^{n} (c_a - c_a')P_a^{(1)}((\nu)) = 0,$$

for all "admissible" values of the variables These values, say $Q_a^{(1)}$, are precisely those which satisfy the conditions (3).

It is evidently easy to construct a system of $Q_a^{(1)}$, $1 \leqq a \leqq n$, which satisfy the conditions (3) written with the sign of strict inequality ">". Hence in a sufficiently small neighborhood of the point $(Q_1^{(1)}, Q_2^{(1)}, \cdots, Q_n^{(1)})$ in the n-dimensional space these strict inequalities still hold. Hence the polynomial vanishes in this neighborhood and so must vanish identically; that is,

$$c_a - c_a' = 0 \quad \text{for} \quad 0 \leqq a \leqq n. \qquad \text{Q. E. D.}$$

ON THE EFFICIENT DESIGN OF STATISTICAL INVESTIGATIONS

By Abraham Wald

Columbia University

1. Introduction. A theory of efficient design of statistical investigations has been developed by R. A. Fisher[1] and his followers mainly in connection with agricultural experimentation. However, the same methods can be applied to other fields also. All statistical designs treated in the aforementioned theory refer to problems of testing linear hypotheses. By testing a linear hypothesis we mean the following problem: Let $y_1, \cdots, y_N$ be N independently and normally distributed variates with a common variance σ^2. It is assumed that the expected value of y_α is given by

$$E(y_\alpha) = \beta_1 x_{1\alpha} + \beta_2 x_{2\alpha} + \cdots + \beta_p x_{p\alpha} \qquad (\alpha = 1, \cdots, N) \tag{1}$$

where the quantities $x_{i\alpha} (i = 1, \cdots, p; \alpha = 1, \cdots, N)$ are known constants and $\beta_1, \cdots, \beta_p$ are unknown constants. The coefficients $\beta_1, \cdots, \beta_p$ are called the population regression coefficients of y on $x_1, x_2, \cdots$, and x_p, respectively. The hypothesis that the unknown regression coefficients $\beta_1, \cdots, \beta_p$ satisfy a set of linear equations

$$g_{i1}\beta_1 + \cdots + g_{ip}\beta_p = g_i \qquad (i = 1, \cdots, r; r \leqq p), \tag{2}$$

is called a linear hypothesis The problem under consideration is that of testing the hypothesis (2) on the basis of the observed values $y_1, \cdots, y_N$.

In many cases the experimenter has a certain amount of freedom in the choice of the values $x_{i\alpha}$. The efficiency of the test is greatly affected by the values of $x_{i\alpha}$. The statistical investigation is efficiently designed if the values $x_{i\alpha}$ are chosen so that the sensitivity of the test is maximized. Let us illustrate this by a simple example. Suppose that x and y have a bivariate normal distribution and we want to test the hypothesis that the regression coefficient β of y on x has a particular value β_0. Suppose, furthermore, that the test has to be carried out on the basis of N pairs of observations $(x_1, y_1), \cdots, (x_N, y_N)$, where the experiments are performed in such a way that $x_1, \cdots, x_N$ are not random variables but have predetermined fixed values. It is known that the variance of the least square estimate b of β is inversely proportional to $\sum_{\alpha=1}^{N} (x_\alpha - \bar{x})^2$ where $\bar{x} = (x_1 + \cdots + x_N)/N$. Hence, if we can freely choose the values $x_1, \cdots, x_N$ in a certain domain D, the greatest sensitivity of the test will be achieved by choosing $x_1, \cdots, x_N$ so that $\Sigma(x_\alpha - \bar{x})^2$ becomes a maximum.

In the next section we will introduce a measure of the efficiency of the design

[1] See for instance R. A. Fisher, *The Design of Experiments*, Oliver and Boyd, London, 1935.

of a statistical investigation for testing a linear hypothesis. In sections 3 and 4 it will be shown that some well known experimental designs, used widely in agricultural experimentation, are most efficient in the sense of the definition given in section 2.

2. A measure of the efficiency of the design of a statistical investigation for testing a linear hypothesis. The hypothesis (2) can be reduced by a suitable linear transformation to the canonical form

$$\beta_1 = \beta_2 = \cdots = \beta_r = 0, \qquad (r \leq p). \tag{3}$$

Hence, we can restrict ourselves without loss of generality to the consideration of the hypothesis (3).

Denote $\sum_{\alpha=1}^{N} x_{i\alpha}x_{j\alpha}$ by a_{ij} and let the matrix $\| c_{ij} \|$ be the inverse of the matrix $\| a_{ij} \|$ $(i, j = 1, \cdots, p)$. Denote by b_i the least square estimate of β_i $(i = 1, \cdots, p)$. It is known that the estimates $b_1, \cdots, b_p$ have a joint normal distribution with mean values $\beta_1, \cdots, \beta_p$, respectively. It is furthermore known that the covariance of b_i and b_j is equal to $c_{ij}\sigma^2$. The statistic used for testing the hypothesis (3) is given by

$$F = \frac{N - p}{r} \frac{\sum_{l=1}^{r}\sum_{m=1}^{r} a^*_{lm} b_l b_m}{\sum_{\alpha=1}^{N} (y_\alpha - b_1 x_{1\alpha} \cdots - b_p x_{p\alpha})^2} \tag{4}$$

where $\| a^*_{lm} \|$ is the inverse of $\| c_{lm} \|$ $((l, m = 1, \cdots, r)$. The statistic F has the F-distribution with r and $N - p$ degrees of freedom. The critical region for testing the hypothesis (3) is given by the inequality

$$F \geqq F_0, \tag{5}$$

where the constant F_0 is determined so that the probability that $F \geq F_0$ (calculated under the assumption that (3) holds) is equal to the level of significance we wish to have

It is known that the power function[2] of the critical region (5) depends only on the single parameter

$$\lambda = \frac{1}{\sigma^2} \sum_{l=1}^{r}\sum_{m=1}^{r} a^*_{lm}\beta_l\beta_m. \tag{6}$$

Furthermore this power function is a monotonically increasing function of λ. The coefficients a^*_{lm} are functions of the quantities $x_{i\alpha}$ $(i = 1, \cdots, p;\ \alpha = 1, \cdots, N)$. The choice of the values $x_{i\alpha}$ $(i = 1, \cdots, p;\ \alpha = 1, \cdots, N)$ is the better the greater the corresponding value of λ. If $r = 1$, the expression λ

[2] See for instance P. C. TANG, "The power function of the analysis of variance tests," *Stat. Res Mem.*, Vol. II, 1938.

reduces to $\frac{1}{\sigma^2} a_{11}^* \beta_1^2$. Hence, if $r = 1$, we maximize λ by maximizing a_{11}^*. Since $a_{11}^* = 1/c_{11}$, we maximize λ by minimizing c_{11}. Thus, if $r = 1$, we can say that we obtain the most powerful test by minimizing c_{11}, i.e. by minimizing the variance of b_1. If $r > 1$, the difficulty arises that no set of values $x_{i\alpha}$ $(i = 1, \cdots, p; \alpha = 1, \cdots, N)$ can be found for which λ becomes a maximum irrespective of the values of the unknown parameters $\beta_1, \cdots, \beta_r$. Hence, if $r > 1$, we have to be satisfied with some compromise solution. For this purpose let us consider the unit sphere

$$\beta_1^2 + \cdots + \beta_r^2 = 1, \tag{7}$$

in the space of the parameters $\beta_1, \cdots, \beta_r$. It is known that the smallest root in ρ of the determinantal equation

$$\begin{vmatrix} a_{11}^* - \rho & a_{12}^* & \cdots & a_{1r}^* \\ a_{21}^* & a_{22}^* - \rho & \cdots & a_{2r}^* \\ \cdot & \cdot & \cdot & \cdot \\ a_{r1}^* & a_{r2}^* & \cdots & a_{rr}^* - \rho \end{vmatrix} = 0, \tag{8}$$

is equal to the minimum value of $\sigma^2\lambda$ on the unit sphere (7). Similarly the greatest root of (8) is equal to the maximum value of $\sigma^2\lambda$ on the sphere (7). The compromise solution of maximizing the smallest root of (8) seems to be a very reasonable one. However, for the sake of certain mathematical simplifications, we propose to maximize the product of the r roots of (8). Since the product of the roots of (8) is equal to the determinant

$$\begin{vmatrix} a_{11}^* \cdots a_{1r}^* \\ \cdot \cdots \cdot \\ a_{r1}^* \cdots a_{rr}^* \end{vmatrix}, \tag{9}$$

we have to maximize the determinant (9). The value of the determinant $|c_{lm}|$ $(l, m = 1, \cdots, r)$ is the reciprocal of that of (9). Hence we maximize (9) by minimizing the determinant $|c_{lm}|$. The generalized variance of the set of variates $b_1, \cdots, b_r$ is equal to the product of σ^{2r} and the determinant $|c_{lm}|$. Thus, our result can be expressed as follows: The optimum choice of the values of $x_{i\alpha}$ is that for which the generalized variance of the variates $b_1, \cdots, b_r$ becomes a minimum.

Any set of pN values $x_{i\alpha}$ $(i = 1, \cdots, p; \alpha = 1, \cdots, N)$ can be represented by a point in the pN-dimensional Cartesian space. Denote by D the set of all points in the pN-dimensional space which we are free to choose. If N is fixed and if any point of D can be equally well chosen, the following two definitions seem to be appropriate:

DEFINITION 1. *Denote by c the minimum value of the determinant $|c_{lm}|$ $(l, m = 1, \cdots, r)$ in the domain D. Then the ratio $c/|c_{lm}|$ is called the efficiency of the design of the statistical investigation for testing the hypothesis* (3).

DEFINITION 2. *The design of the statistical investigation for testing the hypothesis* (3) *is said to be most efficient if its efficiency is equal to* 1.

3. Efficiency of the Latin square design. A widely used and important design in agricultural experimentation is the so-called Latin square. Suppose we wish to find out by experimentation whether there is any significant difference among the yields of m different varieties $v_1, \cdots, v_m$. For this purpose the experimental area is subdivided into m^2 plots lying in m rows and m columns and each plot is assigned to one of the varieties $v_1, \cdots, v_m$. If each variety appears exactly once in each row and exactly once in each column, we have a Latin square arrangement. Denote by y_{ijk} the yield of the variety v_k on the plot which lies in the i-th row and j-th column. The subscript k is, of course, a single valued function of the subscripts i and j, since to each plot only one variety is assigned. The following assumptions are made: the variates y_{ijk} are independently and normally distributed with a common variance σ^2 and the expected value of y_{ijk} is given by

$$E(y_{ijk}) = \mu_i + \nu_j + \rho_k. \tag{10}$$

The parameters σ^2, μ_i, ν_j and ρ_k are unknown. The hypothesis to be tested is the hypothesis that variety has no effect on yield, i.e.

$$\rho_1 = \rho_2 = \cdots = \rho_k. \tag{11}$$

We associate the positive integer $\alpha(i, j) = (i - 1)m + j$ with the plot which lies in the i-th row and j-th column $(i, j = 1, \cdots, m)$. It is clear that for any positive integer $\alpha \leq m^2$ there exists exactly one plot, i.e. exactly one pair of values i and j, such that $\alpha = \alpha(i, j)$. In the following discussions the symbol y_α $(\alpha = 1, \cdots, m^2)$ will denote the yield y_{ijk} where the indices i and j are determined so that $\alpha(i, j) = \alpha$. The plot in the i-th row and j-th column will be called the α-th plot where $\alpha = \alpha(i, j)$.

We define the symbols $t_{i\alpha}$, $u_{j\alpha}$, $z_{k\alpha}$ $(i, j, k = 1, \cdots, m;\ \alpha = 1, \cdots, m^2)$, as follows: $t_{i\alpha} = 1$ if the α-th plot lies in the i-th row, and $t_{i\alpha} = 0$ otherwise. Similarly $u_{j\alpha} = 1$ if the α-th plot lies in the j-th column, and $u_{j\alpha} = 0$ otherwise. Finally $z_{k\alpha} = 1$ if the k-th variety is assigned to the α-th plot, and $z_{k\alpha} = 0$ otherwise. Then equation (10) can be written as

$$\begin{aligned} E(y_\alpha) = {} & \mu_1 t_{1\alpha} + \cdots + \mu_m t_{m\alpha} + \nu_1 u_{1\alpha} + \cdots \\ & + \nu_m u_{m\alpha} + \rho_1 z_{1\alpha} + \cdots + \rho_m z_{m\alpha}. \end{aligned} \tag{12}$$

Denote the arithmetic means $\frac{1}{m^2}\sum_{\alpha=1}^{m^2} t_{i\alpha}$, $\frac{1}{m^2}\sum_{\alpha=1}^{m^2} u_{i\alpha}$, and $\frac{1}{m^2}\sum_{\alpha=1}^{m^2} z_{i\alpha}$ by $\bar{t}_i$, $\bar{u}_i$ and $\bar{z}_i$ respectively. Let $t'_{i\alpha} = t_{i\alpha} - \bar{t}_i$, $u'_{i\alpha} = u_{i\alpha} - \bar{u}_i$, $z'_{i\alpha} = z_{i\alpha} - \bar{z}_i$, $\mu'_i = \mu_i - \mu_m$, $\nu'_i = \nu_i - \nu_m$ and $\rho'_i = \rho_i - \rho_m$ for $i = 1, \cdots, m - 1$. Let furthermore $w_\alpha = 1$ for $\alpha = 1, \cdots, m^2$. Then we have

$$\begin{cases} t_{i\alpha} = t'_{i\alpha} + \bar{t}_i w_\alpha; \quad u_{i\alpha} = u'_{i\alpha} + \bar{u}_i w_\alpha; \quad z_{i\alpha} = z'_{i\alpha} + \bar{z}_i w_\alpha; \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (i = 1, \cdots, m - 1) \\ t_{m\alpha} = (1 - \bar{t}_1 - \cdots - \bar{t}_{m-1})w_\alpha - t'_{1\alpha} - \cdots - t'_{m-1,\alpha}, \\ u_{m\alpha} = (1 - \bar{u}_1 - \cdots - \bar{u}_{m-1})w_\alpha - u'_{1\alpha} - \cdots - u'_{m-1,\alpha}, \\ z_{m\alpha} = (1 - \bar{z}_1 - \cdots - \bar{z}_{m-1})w_\alpha - z'_{1\alpha} - \cdots - z'_{m-1,\alpha}. \end{cases} \tag{13}$$

From (12) and (13) we obtain

$$E(y_\alpha) = \xi w_\alpha + \sum_{i=1}^{m-1} \mu_i' t_{i\alpha}' + \sum_{i=1}^{m-1} \nu_i' u_{i\alpha}' + \sum_{i=1}^{m-1} \rho_i' z_{i\alpha}' \tag{14}$$

where

$$\xi = \sum_{i=1}^{m-1} \mu_i' \bar{t}_i + \sum_{i=1}^{m-1} \nu_i' \bar{u}_i + \sum_{i=1}^{m-1} \rho_i' \bar{z}_i + \mu_m + \nu_m + \rho_m .$$

The hypothesis (11) can be written as

$$\rho_1' = \rho_2' = \cdots = \rho_{m-1}' = 0. \tag{15}$$

This is a linear hypothesis in canonical form as given in (3). The values $z_{i\alpha}'$ $(i = 1, \cdots, m-1;\ \alpha = 1, \cdots, m^2)$ depend on the way in which the varieties $v_1, \cdots, v_m$ are assigned to the m^2 plots. We will show that we obtain a most efficient design if we distribute the varieties over the m^2 plots in a Latin square arrangement, i e. if each variety appears exactly once in each row and exactly once in each column.

Let $q_{1\alpha} = w_\alpha$, $q_{i+1,\alpha} = t_{i\alpha}'$ $(i = 1, \cdots, m-1)$, $q_{m+j,\alpha} = u_{j\alpha}'$ $(j = 1, \cdots, m-1)$ and $q_{2m-1+k,\alpha} = z_{k\alpha}'$ $(k = 1, \cdots, m-1)$. Denote $\sum_{\alpha=1}^{m^2} q_{i\alpha} q_{j\alpha}$ by a_{ij} $(i, j = 1, 2, \cdots, 3m-2)$ and let the matrix $\| c_{ij} \|$ be the inverse of the matrix $\| a_{ij} \|$ $(i, j = 1, \cdots, 3m-2)$. Let us denote by Δ the determinant $| a_{ij} |$ $(i, j = 1, \cdots, 3m-2)$, by Δ_1 the determinant $| a_{ij} |$ $(i, j = 1, \cdots, 2m-1)$, by Δ_2 the determinant $| a_{ij} |$ $(i, j = 2m, \cdots, 3m-2)$ and Δ_2' the determinant $| c_{ij} |$ $(i, j = 2m, \cdots, 3m-2)$. We have to show that for the Latin square arrangement Δ_2' becomes a minimum. From a known theorem[3] about determinants it follows that

$$\Delta_2' = \Delta_1 / \Delta. \tag{16}$$

Hence, we have merely to show that Δ/Δ_1 becomes a maximum for the Latin square arrangement. Denote by $\bar{\Delta}$, $\bar{\Delta}_1$ and $\bar{\Delta}_2$ the values taken by Δ, Δ_1 and Δ_2, respectively, in the case of a Latin square arrangement. Since, for the Latin square arrangement, as is known,

$$\sum_{\alpha=1}^{m^2} z_{k\alpha}' u_{j\alpha}' = \sum_{\alpha=1}^{m^2} z_{k\alpha}' t_{i\alpha}' = \sum_{\alpha=1}^{m^2} z_{k\alpha}' w_\alpha = 0 \qquad (i, j, k = 1, \cdots, m-1)$$

we have

$$\frac{\bar{\Delta}}{\bar{\Delta}_1} = \bar{\Delta}_2 . \tag{17}$$

Since the matrix $\| a_{ij} \|$ $(i, j = 1, \cdots, 3m-2)$ is positive definite we have

$$\frac{\Delta}{\Delta_1} \leqq \Delta_2 . \tag{18}$$

[3] See M. Bôcher, *Introduction to Higher Algebra*, 1931, pp. 31.

Because of (17) and (18) the Latin square design is proved to be most efficient if we show that $\Delta_2 \leq \bar{\Delta}_2$.

Denote by Δ_2^* the m-rowed determinant $|a_{ij}|$ $(i, j = 1, 2m, 2m+1, \cdots, 3m-2)$. Since $a_{1j} = 0$ for $j \neq 1$, we have

$$\Delta_2^* = a_{11}\Delta_2 = m^2\Delta_2. \tag{19}$$

Denote $\sum_{\alpha=1}^{m^2} z_{i\alpha}z_{j\alpha}$ by b_{ij} $(i, j = 1, \cdots, m)$. Then

$$\begin{cases} b_{ij} = 0, & \text{for } i \neq j \\ \text{and } b_{ii} = N_i, \end{cases} \tag{20}$$

where N_i denotes the number of plots to which the variety v_i has been assigned. Because of (20) we have

$$\begin{vmatrix} b_{11} & \cdots & b_{1m} \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ b_{m1} & & b_{mm} \end{vmatrix} = N_1N_2 \cdots N_m. \tag{21}$$

According to (13) we have

$$\begin{aligned} z'_{i\alpha} + \bar{z}_i w_\alpha &= z_{i\alpha}, \qquad (i = 1, \cdots, m-1) \\ -z'_{1\alpha} - \cdots - z'_{m-1,\alpha} + w_\alpha(1 - \bar{z}_1 - \cdots - \bar{z}_{m-1}) &= z_{m\alpha}. \end{aligned} \tag{22}$$

The determinant of these equations is given by

$$\lambda = \begin{vmatrix} 1 & 0 & 0 & \cdots & 0 & 0 & \bar{z}_1 \\ 0 & 1 & 0 & \cdots & 0 & 0 & \bar{z}_2 \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdots & 0 & 1 & \bar{z}_{m-1} \\ -1 & -1 & -1 & \cdots & -1 & -1 & \delta \end{vmatrix} \tag{23}$$

where $\delta = 1 - \bar{z}_1 - \bar{z}_2 - \cdots - \bar{z}_{m-1}$. It is easy to verify that

$$\lambda = 1. \tag{24}$$

From (21), (22) and (24) it follows that

$$\Delta_2^* = N_1N_2 \cdots N_m. \tag{25}$$

Hence, from (19) we obtain

$$\Delta_2 = N_1N_2 \cdots N_m/m^2 \tag{26}$$

In the case of a Latin square design we have $N_1 = N_2 = \cdots = N_m = m$. Hence

$$\bar{\Delta}_2 = m^{m-2}. \tag{27}$$

Because of the condition $N_1 + N_2 + \cdots + N_m = m^2$, the right hand side of (26) becomes a maximum when $N_1 = N_2 = \cdots = N_m = m$. Thus $\Delta_2 \leq \bar{\Delta}_2$ and consequently the Latin square design is proved to be most efficient.

4. Efficiency of Graeco-Latin and higher squares. Consider m varieties $v_1, \cdots, v_m$ and m treatments $q_1, \cdots, q_m$ Suppose that we wish to find out by experimentation whether the yield is affected by varieties or treatments. For this purpose the experimental area is subdivided into m^2 plots lying in m rows and m columns and to each plot one of the varieties and one of the treatments is assigned. We call this arrangement a Graeco-Latin square if the following conditions are fulfilled: 1) each variety appears exactly once in each row and exactly once in each column; 2) each treatment appears exactly once in each row and exactly once in each column; 3) each variety is combined with each of the treatments exactly once.

The following general abstract scheme includes the Latin square and Graeco-Latin square as special cases: Consider an r-way classification with m classes in each classification. Denote by $y_{a_1 a_2 \cdots a_r}$ the value of a certain characteristic of an individual who is classified in the a_1-class of the first classification, in the a_2-class of the second classification, $\cdots$, and in the a_r-class of the r-th classification. Suppose that m^2 observations are made for the purpose of investigating the effect of the classes on the value of the characteristic under consideration. We will say that we have a generalized Latin square design if the following condition is fulfilled: *Let i, j, m' and m'' be an arbitrary set of four positive integers for which $i \neq j$, $i \leq r$, $j \leq r$, $m' \leq m$ and $m'' \leq m$. Then among the m^2 individuals observed there exists exactly one individual who belongs to the m'-class of the i-th classification and m''-class of the j-th classification.*

It is clear that if $r = 3$ the above scheme is a Latin square. If $r = 4$ we have a Graeco-Latin square.

Assume that the observations $y_{a_1 \cdots a_r} (a_1, a_2, \cdots, a_r = 1, \cdots, m)$ are normally and independently distributed with a common variance σ^2. Assume furthermore that the expected value of $y_{a_1 \cdots a_r}$ is given by

$$E(y_{a_1 a_2 \cdots a_r}) = \gamma_{1a_1} + \cdots + \gamma_{ra_r}.$$

The parameters σ^2 and γ_{ia} $(i = 1, \cdots, r;\ a = 1, \cdots, m)$ are unknown constants. Suppose that we wish to test the hypothesis that

$$\gamma_{i1} = \gamma_{i2} = \cdots = \gamma_{im}. \tag{28}$$

It can be shown that if the number of observations is limited to m^2, we obtain a most efficient design by constructing a generalized Latin square. The proof of this statement is similar to that of the efficiency of the Latin square and is therefore omitted.

SOME SIGNIFICANCE TESTS FOR NORMAL BIVARIATE DISTRIBUTIONS

By D. S. Villars and T. W. Anderson

United States Rubber Company, Passaic, New Jersey, and Princeton University

1. Introduction. In the theory of linear regression of y on x where y is normally distributed about a linear function of x, say $\nu + \beta x$, where x is a "fixed" variate, the t-test for the hypothesis that β is zero (that y is distributed about ν; independent of x) is well known. In this paper we apply some general statistical theory to the similar problem where x and y are jointly normally distributed. This case is commonly known as the case of "error in both variates." We derive a criterion for testing the hypothesis that the population means are the coordinates of a specified point when the ratio of the variances and the population correlation coefficient are known When the ratio of variances is known, a criterion is derived to test whether the correlation coefficient is zero.

2. The means. Let us consider a sample of n pairs of observations $(x_1, y_1; x_2, y_2; \cdots; x_n, y_n)$ from a normal bivariate population. Let the variances of x and of y be σ_x^2 and σ_y^2, respectively; and the correlation coefficient, say ρ, be zero. Suppose the ratio of the weight of y to the weight of x, say $\gamma = w_y/w_x = \sigma_x^2/\sigma_y^2$, is known although the variances are not known. It is clear then, that $\sqrt{\gamma}\, y$ has variance σ_x^2. Since the observations y_i $(i = 1, 2, \cdots, n)$ can be transformed into revised observations $\sqrt{\gamma}\, y_i = y_i'$, we lose no generality by assuming that x and y are both distributed with variance σ^2.

Under the assumption of equality of variances and independence of variates we shall derive a criterion for testing the null hypothesis that each observation x_i is of a variate distributed about the same population mean μ and each observation y_i is of a variate distributed about the same population mean ν. The hypothesis may be stated symbolically as:

$$H_0: E(x) = \mu, \qquad E(y) = \nu,$$

given $\sigma_x^2 = \sigma_y^2 = \sigma^2$ and $\rho = 0$. We can write

$$\sum_{i=1}^{n} (x_i - \mu)^2 = n(\bar{x} - \mu)^2 + S_x,$$

$$\sum_{i=1}^{n} (y_i - \nu)^2 = n(\bar{y} - \nu)^2 + S_y,$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

$$S_x = \sum_{i=1}^{n} (x_i - \bar{x})^2, \qquad S_y = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

Then $n(\bar{x} - \mu)^2/\sigma^2$ and $n(\bar{y} - \nu)^2/\sigma^2$ are each distributed independently as χ^2 with one degree of freedom and each of S_x/σ^2 and S_y/σ^2 follow the χ^2-law with $n - 1$ degrees of freedom. If we define

$$r = \sqrt{(\bar{x} - \mu)^2 + (\bar{y} - \nu)^2}, \qquad S_r = S_x + S_y, \tag{1}$$

then nr^2/σ^2 and S_r/σ^2 have independent χ^2-distributions with 2 and $2n - 2$ degrees of freedom, respectively.

It follows from this that

$$R = \frac{nr^2}{2\sigma^2} \Big/ \frac{S_r}{(2n-2)\sigma^2} = n(n-1)\frac{r^2}{S_r} = n(n-1)\frac{(\bar{x} - \mu)^2 + (\bar{y} - \nu)^2}{S_x + S_y}, \tag{2}$$

has the F-distribution with 2 and $2n - 2$ degrees of freedom.

Let us define F_α so

$$\int_{F_\alpha}^{\infty} h_{2,2n-2}(F)\, dF = \alpha, \tag{3}$$

where $h_{2,2n-2}(F)$ is the F-distribution with 2 and $2n - 2$ degrees of freedom and $0 \leq \alpha \leq 1$. Then the probability is α that the sample statistic R is greater than or equal to F_α, i.e.,

$$P\{R \geq F_\alpha\} = \alpha. \tag{4}$$

In considering a sample value of R, at significance level α, one rejects the hypothesis of the means being μ and ν, respectively, if R is larger than F_α, i.e., larger than 1 and larger than the α significance point in Snedecor's tables [1].

This F-test is a straightforward generalization to the bivariate case of the usual t-test as applied to the univariate case. In each case the sum of squares of distances of the observations from the population mean is broken up into the sum of squares of distances from the sample mean plus n times the square of the distance from the sample mean to the population mean. The t-test for the univariate case depends on the ratio of the distance of the sample mean from the population mean to the square root of the sum of squares of distances from the observations to the sample mean. The proposed F-test depends upon the ratio of the square of the distance of the sample mean from the population mean to the sum of squares of distances from the observations to the sample mean.

It can easily be shown that the likelihood ratio criterion for this hypothesis is

$$\lambda = \left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(y_i - \bar{y})^2}{\sum_{i=1}^{n}(x_i - \mu)^2 + \sum_{i=1}^{n}(y_i - \nu)^2}\right]^n = \left[1 + \frac{R}{n-1}\right]^n. \tag{5}$$

The hypothesis considered here is one of a class of hypotheses treated by Kolodziejczyk [2] in a paper in which he considers the likelihood ratio criterion for a set of general linear hypotheses.

Equation (4) may be written

$$P\{(\bar{x} - \mu)^2 + (\bar{y} - \nu)^2 \geq r_\alpha^2\} = \alpha, \tag{6}$$

where $r_\alpha^2 = F_\alpha (S_x + S_y)/[n(n-1)]$. The probability is α that the distance from the sample means $\bar{x}, \bar{y}$ to the population means μ, ν is greater than or equal to r_α. We may call r_α the *fiducial radius* [3], and the equation $(\bar{x} - \mu)^2 + (\bar{y} - \nu)^2 = r_\alpha^2$ defines the *confidence region* for the population means.

Suppose we have two samples of n_1 and n_2 pairs of observations, respectively, from normal bivariate distributions. If the population mean of each x variate is μ and the population mean of each y variate is ν, the population variance of each variate is σ^2, and the correlation coefficient is zero, then the sample means $\bar{x}_1$ and $\bar{y}_1$ of the first sample and $\bar{x}_2$ and $\bar{y}_2$ of the second sample follow normal distributions. Also $\bar{x}_1 - \bar{x}_2$ and $\bar{y}_1 - \bar{y}_2$ are normally distributed. Then $r'^2 = n_1 n_2/(n_1 + n_2)[(\bar{x}_1 - \bar{x}_2)^2 + (\bar{y}_1 - \bar{y}_2)^2]/\sigma^2$ has the χ^2-distribution with 2 degrees of freedom. Let

$$S'_{r'} = \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2 + \sum_{i=1}^{n_1} (y_{1i} - \bar{y}_1)^2 + \sum_{i=1}^{n_2} (x_{2i} - \bar{x}_2)^2 + \sum_{i=1}^{n_2} (y_{2i} - \bar{y}_2)^2,$$

where x_{1i}, y_{1i} $(i = 1, 2, \cdots, n_1)$ are the pairs of observations in the first sample and x_{2i}, y_{2i} $(i = 1, 2, \cdots, n_2)$ are the pairs of observations in the second sample. $S'_{r'}/\sigma^2$ is distributed according to the χ^2-distribution with $(2n_1 + 2n_2 - 4)$ degrees of freedom because it is the sum of quantities independently distributed as χ^2. Then

$$R' = \frac{n_1 n_2 r'^2}{2(n_1 + n_2)\sigma^2} \Big/ \frac{S'_{r'}}{(2n_1 + 2n_2 - 4)\sigma^2} = \frac{n_1 n_2 (n_1 + n_2 - 2) r'^2}{(n_1 + n_2) S'_{r'}}$$

has the F-distribution with 2 and $(2n_1 + 2n_2 - 4)$ degrees of freedom. This fact yields us a significance test for the hypothesis that both the means of the x variates and the means of the y variates for the two populations are the same. We can also set up confidence regions for $\mu_1 - \mu_2$ and $\nu_1 - \nu_2$.

Now let us consider a sample from a normal bivariate population with means μ and ν, variances σ_x^2 and σ_y^2 and correlation coefficient ρ. Suppose $\gamma = \sigma_x^2/\sigma_y^2$ and ρ are known. The transformation

$$x = \frac{\sqrt{1+\rho}\, x' + \sqrt{1-\rho}\, y'}{\sqrt{2}}, \qquad y = \frac{\sqrt{1+\rho}\, x' - \sqrt{1-\rho}\, y'}{\sqrt{2}\,\gamma}, \tag{8}$$

gives us the variates x' and y' which are distributed independently and with variance σ_x^2. Applying the results above we see that

$$\begin{aligned} R &= n(n-1) \frac{(\bar{x}' - \mu')^2 + (\bar{y}' - \nu')^2}{\sum_{i=1}^{n} (x'_i - \bar{x}')^2 + \sum_{i=1}^{n} (y'_i - \bar{y}')^2} \\ &= n(n-1) \frac{(\bar{x} - \mu)^2 - 2\rho\sqrt{\gamma}\,(\bar{x} - \mu)(\bar{y} - \nu) + \gamma(\bar{y} - \nu)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2 - 2\rho\sqrt{\gamma} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) + \gamma \sum_{i=1}^{n} (y_i - \bar{y})^2} \end{aligned} \tag{9}$$

has the F-distribution with 2 and $2n - 2$ degrees of freedom. From this we derive significance tests, fiducial radii, and confidence regions as before.

The above distributions, significance tests, and confidence regions are easily generalized to multivariate normal distributions. Suppose we have a sample of n k-tuples of observations $\{x_{i\alpha}\}$ $(i = 1, 2, \cdots, k; \alpha = 1, 2, \cdots, n)$ from a k-variate normal distribution. Let the expected value of each variate x_i be zero $(i = 1, 2, \cdots, k)$, the variance of each variate be σ^2 and each correlation coefficient be zero. Then

$$R'' = \frac{n(n-1)\sum_{i=1}^{k} \bar{x}_i^2}{\sum_{i=1}^{k}\sum_{\alpha=1}^{n} (x_{i\alpha} - \bar{x}_i)^2} \tag{10}$$

has the F-distribution with k and $k(n-1)$ degrees of freedom. Significance tests, confidence regions, and fiducial radii follow from this fact.

3. Linear Regression. If one has a sample of n pairs of observations $(x_1, y_1; x_2, y_2; \cdots; x_n, y_n)$ from a normal bivariate population and wishes to fit a straight line to the scatter of sample points, one fits the line in such a way that the sum of squares of distances from the sample points to the line is a minimum ("error in both variates").

It is easily shown that this line goes through the point whose coordinates are the sample means $(\bar{x}, \bar{y})$. If the slope of a line through $(\bar{x}, \bar{y})$ is $\tan\theta$, the distance from a sample point (x_i, y_i) to the line is $(x_i - \bar{x}) \sin\theta - (y_i - \bar{y}) \cos\theta$. The sum of squares of distances from sample points to the line is

$$\sin^2\theta\, S_x - 2 \sin\theta \cos\theta\, S_{xy} + \cos^2\theta\, S_y,$$

where

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

If we minimize the above expression with respect to θ we find

$$b = \tan\theta = \frac{S_y - S_x \pm \sqrt{(S_y - S_x)^2 + 4S_{xy}^2}}{2S_{xy}}. \tag{11}$$

Using the plus sign gives us S_p, the minimum sum of squared distances; using the minus sign gives us S_a, the maximum sum of squared distances. (The latter value of $\tan\theta$ is the negative reciprocal of the former.)

S_p is the sum of squared distances perpendicular to the regression line and S_a is the sum of squared distances along the regression line. The sum $S_p + S_a$ is equal to $S_x + S_y$ which is the sum of squares of distances from the sample points to the point $\bar{x}, \bar{y}$. We have thus decomposed $S_x + S_y$ into two components, one perpendicular to the regression line and the other along the regression line.

The joint distribution of S_p and S_a may be derived from the Wishart distribution of the sums of squares and cross products,[1]

$$(12) \qquad \frac{1}{4\pi\sigma^{2n-2}\Gamma(n-2)} \begin{vmatrix} S_x & S_{xy} \\ S_{xy} & S_y \end{vmatrix}^{\frac{1}{2}(n-4)} e^{-\frac{1}{2}(S_x+S_y)/\sigma^2}$$

Let us make the transformation

$$S_x = \cos^2\theta\, S_a + \sin^2\theta\, S_p\,,$$

$$S_y = \sin^2\theta\, S_a + \cos^2\theta\, S_p\,,$$

$$S_{xy} = \sin\theta\cos\theta\,(S_a - S_p).$$

The value of θ corresponds to the plus sign in (11). We find

$$S_x + S_y = S_p + S_a\,,$$

$$\begin{vmatrix} S_x & S_{xy} \\ S_{xy} & S_y \end{vmatrix} = S_p S_a\,.$$

The Jacobian of the transformation is $(S_a - S_p)$. Using these relations in (12) and integrating out θ we derive the distribution of S_a and S_p

$$(13) \qquad \frac{1}{4\sigma^4\Gamma(n-2)} \left(\frac{S_a S_p}{\sigma^4}\right)^{\frac{1}{2}(n-4)} e^{-\frac{1}{2}(S_a+S_p)/\sigma^2}\,(S_a - S_p).$$

It can be shown that S_a and S_p are the characteristic roots of the sample variance-covariance matrix. The distribution (13) of the characteristic roots of a variance-covariance matrix when the population correlation coefficient is zero and the variances are equal has been demonstrated by P. L. Hsu [4]

As a test of correlation (i.e., test of significance of the regression coefficient) we propose using the ratio

$$F' = S_a/S_p\,.$$

This ratio is the maximum ratio of the sum of squared deviations in one direction to the sum of squared deviations in the perpendicular direction. It is intuitively evident that this ratio is probably near unity if the null hypothesis is true, that is, if the variances are equal and the correlation is zero. If the correlation is not zero then the ratio is likely to be large.

From (13) we can deduce the distribution of F' by transforming variables and integrating out the extraneous one. This procedure yields us as the distribution of F'

$$(n-2)2^{n-3}F'^{\frac{1}{2}(n-4)}(F'+1)^{-(n-1)}(F'-1).$$

If we make the transformation

$$F' = e^{2z'},$$

[1] This distribution is equivalent to Fisher's distribution of the sample variances and correlation coefficient when the population correlation coefficient is zero.

we find the probability element of z' to be

$$(n - 2)(\cosh z')^{-(n-1)}\, d(\cosh z')$$

After integrating we see the cumulative distribution of z' is

$$1 - (\cosh z')^{-(n-2)}.$$

Critical values of z' for various levels of significance may be determined from a table of hyperbolic cosines. Table I gives some values of z' and the corresponding values of F'.

TABLE I

Percentage points for the z' (or F') distribution

n	z'					F'				
	$P_{.20}$	$P_{.10}$	$P_{.05}$	$P_{.01}$	$P_{.001}$	$P_{.20}$	$P_{.10}$	$P_{.05}$	$P_{.01}$	$P_{.001}$
3	2 292	2 993	3.688	5.298	7.601	98.0	398	1600	40,000	4,000,000
4	1.444	1.818	2 178	2.993	4.144	17.9	38 0	78.0	398	4,000
5	1 130	1 402	1.656	2.216	2 993	9 59	16.5	27 4	84.2	398
6	.958	1.178	1 381	1.818	2.412	6.79	10.6	15.8	38.0	124
7	846	1.035	1.207	1.572	2.059	5 43	7 92	11.2	23 2	61.4
8	766	933	1 084	1 402	1.818	4.63	6.47	8 74	16.5	38.0
9	704	856	992	1.276	1 643	4.09	5 55	7.28	12.7	26.8
10	656	.796	.920	1 178	1.509	3.71	4.91	6 30	10.6	20 5
11	.616	746	.862	1.100	1 402	3 43	4.45	5 61	9.02	16.5
12	.583	.705	.813	1.035	1.314	3.21	4.10	5.09	7.92	13.9
13	554	.670	.772	980	1.241	3.03	3 82	4.68	7.10	12.0
14	.530	.639	736	.933	1 178	2.89	3.59	4.36	6.47	10.6
15	.508	.613	.705	.892	1 124	2.76	3 41	4.10	6.00	9.47
20	.429	517	.593	.746	993	2.36	2.81	3.27	4.45	6.47
25	378	.455	.522	.654	814	2.13	2.48	2 84	3.70	5.10
30	.342	411	471	.589	732	1.98	2 28	2.57	3.25	4 32
40	293	.352	.402	502	621	1.80	2 02	2.23	2.73	3 47
60	.237	284	324	404	.498	1.61	1.76	1 91	2.24	2.71
120	.165	.198	226	281	.345	1 39	1.49	1.57	1.75	2.00

The use of F' has been suggested here to test the hypothesis that the population correlation coefficient is zero when it is known that the variances of the two variates are the same, or, more generally, when the ratio of the two variances is known. This gives a test of significance of the regression coefficient when there is error in both variates if the ratio of the variances is known. The test arises from intuitive considerations. F' can also be used to test the hypothesis that $\rho = 0$ *and* $\sigma_x^2 = \sigma_y^2$ (H_4 in Hsu's paper). C. T. Hsu [5] and J. W. Mauchly [6] have shown that the likelihood ratio criterion for this hypothesis is

$$\lambda = \left[\frac{2(S_x S_y - S_{xy}^2)}{(S_x + S_y)^2}\right]^{\frac{1}{2}n} = \left[\frac{2F'}{(F' + 1)^2}\right]^{\frac{1}{2}n}.$$

If we set the normal distribution function equal to a constant, we determine a contour ellipse in the x, y — plane. Since these ellipses of constant probability density are circles when $\rho = 0$ and $\sigma_x^2 = \sigma_y^2$, Mauchly calls the test a test of circularity. The same procedure as used to test whether these ellipses are circles can be used to test whether the ellipses have major axes in a certain direction and with a specified ratio of lengths of axes. Suppose we wish to test the hypothesis that the major axis is inclined to the x axis at an angle θ and that the ratio of lengths of the major axis to the minor axis is k. This is equivalent to the hypothesis that $\rho = \rho_0$ and $\sigma_x^2 = \gamma_0\sigma_y^2$. To do this we rotate coordinate axes of the variables of the distribution (hence changing coordinates of all sample points) through θ and change the scale of one of the new variables by the factor of k. The transformation is

$$x = kx' \cos\theta - y' \sin\theta,$$

$$y = kx' \sin\theta + y' \cos\theta.$$

In terms of x', y' the null hypothesis is $\rho' = 0$, $\sigma_{x'}^2 = \sigma_{y'}^2$, and one proceeds as above. Of course, if γ_0 is known then this method can be used to test the null hypothesis that $\rho = \rho_0$.

4. Illustrative Example. An application of the formulae given above may be illustrated from the data in Table II, which gives two sets of electrical conductivity measurements at different field strengths. The assumption that the two variances are equal is thus reasonable.

Table of Pairs of Observations of Electrical Conductivity

x_i	y_i	x_i	y_i
5.0	5.1	5.5	5.1
7.4	7.0	5.3	5.0
7.0	7.7	4.7	4.4
8.8	7.7	8.6	7.1
7.8	6.8	7.5	7.3
5.1	5.5	5.6	6.3
6.6	7.4	7.4	6.5
8.8	7.7		

Is it reasonable to regard x and y as being independently distributed in the population on the basis of these data?

The sums of squares and cross products of deviations from the means and the calculated slope are:

$$S_x = 29.40, \qquad S_{xy} = 19.99,$$

$$S_y = 18.04, \qquad b = 0.7554.$$

The maximized variance ratio is:

$$F' = \frac{S_x + 2bS_{xy} + b^2 S_y}{b^2 S_x - 2bS_{xy} + S_y} = \frac{69.89}{4.615} = 15.15.$$

$$z' = \tfrac{1}{2}\ln F' = 1.36.$$

Comparing with Table I for $n = 15$ we find this value of z' very highly significant (probability less than 0.001), and at this probability level and on basis of our data, x and y cannot be considered to be independent in the population.

Since the regression is significant, it becomes of interest to compute the calculated points X_i and Y_i which fall on the regression line

$$Y = 1.35 + 0.7554\,X,$$

corresponding to each observed point x_i, y_i. They are obtained from these equations

$$Y_i = \bar{y} + \frac{b}{1+b^2}(x_i - \bar{x}) + \frac{b^2}{1+b^2}(y_i - \bar{y})$$

$$= .481x_i + .363y_i + .86,$$

$$X_i = \bar{x} + \frac{1}{1+b^2}(x_i - \bar{x}) + \frac{b}{1+b^2}(y_i - \bar{y})$$

$$= .637x_i + .481y_i - .65.$$

The minimized sum of squared deviations from the regression line (i.e., squared distances between observed and calculated points) is the denominator of the expression for F' divided by the factor $(1 + b^2)$,

$$4.615/.5706 = 2.64.$$

It should perhaps be pointed out that the tests of the means described in the first part of this paper are no longer applicable since we do not know the population correlation coefficient.

REFERENCES

[1] G. W. Snedecor, *Statistical Methods*, Iowa State College Press (1940), pp. 184–187.

[2] S. Kolodziejczyk, "On an important class of statistical hypotheses," *Biometrika*, Vol. 27 (1935), pp. 161–190.

[3] R. A. Fisher, "The fiducial argument in statistical inference," *Annals of Eugenics*, Vol. 6 (1935), pp. 391–398.

[4] P. L. Hsu, "On the distribution of roots of certain determinantal equations," *Annals of Eugenics*, Vol. 9 (1939), pp. 250–258.

[5] C. T. Hsu, "On samples from a normal bivariate population," *Annals of Math. Stat.*, Vol. 11 (1940), pp. 410–426.

[6] J. W. Mauchly, "Significance test for sphericity of a normal n-variate distribution," *Annals of Math. Stat.*, Vol. 11 (1940), pp. 204–209.

SYMMETRIC TESTS OF THE HYPOTHESIS THAT THE MEAN OF ONE NORMAL POPULATION EXCEEDS THAT OF ANOTHER

By Herbert A. Simon

Illinois Institute of Technology

1. Introduction. One of the most commonly recurring statistical problems is to determine, on the basis of statistical evidence, which of two samples, drawn from different universes, came from the universe with the larger mean value of a particular variate. Let M_y be the mean value which would be obtained with universe (Y) and M_x be the mean value which would be obtained with universe (X). Then a test may be constructed for the hypothesis[1] $M_y \geq M_x$.

If $x_1, \cdots, x_n$ are the observed values of the variate obtained from universe (X), and $y_1, \cdots, y_n$ are the observed values obtained from universe (Y), then the sample space of the points $E: (x_1, \cdots, x_n; y_1, \cdots, y_n)$ may be divided into three regions ω_0, ω_1, and ω_2. If the sample point falls in the region ω_0, the hypothesis $M_y \geq M_x$ is accepted; if the sample point falls in the region ω_1, the hypothesis $M_y \geq M_x$ is rejected; if the sample point falls in the region ω_2, judgment is withheld on the hypothesis. Regions ω_0, ω_1, and ω_2 are mutually exclusive and, together, fill the entire sample space. Any such set of regions ω_0, ω_1, and ω_2 defines a test for the hypothesis $M_y \geq M_x$.

In those cases, then, where the experimental results fall in the region ω_2, the test leads to the conclusion that there is need for additional data to establish a result beyond reasonable doubt. Under these conditions, the test does not afford any guide to an unavoidable or non-postponable choice. In the application of statistical findings to practical problems it often happens, however, that judgment can not be held in abeyance—that *some* choice must be made, even at a risk of error. For example, when planting time comes, a choice must be made between varieties (X) and (Y) of grain even if neither has been conclusively demonstrated, up to that time, to yield a larger crop than the other. It is the purpose of this paper to propose a criterion which will always permit a choice between two experimental results, that is, a test in which the regions ω_0 and ω_1 fill the entire sample space. In the absence of a region ω_2, any observed result is interpreted as a definite acceptance or rejection of the hypothesis tested.

2. General characteristics of the criterion. Let us designate the hypothesis $M_y \geq M_x$ as H_0 and the hypothesis $M_x > M_y$ as H_1. Then a pair of tests, T_0 and T_1, for H_0 and H_1 respectively must, to suit our needs, have the following properties:

(1) The regions ω_{00} (ω_{00} is the region of acceptance for H_0, ω_{10} the region of rejection for H_0; ω_{01} and ω_{11} the corresponding regions for H_1) and ω_{11} must

[1] This paper presupposes a familiarity with the theory of testing statistical hypotheses as set forth by J. Neyman and E. S. Pearson [1].

coincide, as must the regions ω_{10} and ω_{01}. This correspondence means that when H_0 is accepted, H_1 is rejected, and vice versa. Hence, the tests T_0 and T_1 are identical, and we shall hereafter refer only to the former.

(2) There must be no regions ω_{20} and ω_{21}. This means that judgment is never held in abeyance, no matter what sample is observed.

(3) The regions ω_{00} and ω_{10} must be so bounded that the probability of accepting H_1 when H_0 is true (error of the first kind for T_0) and the probability of accepting H_0 when H_1 is true (error of the second kind for T_0) are, in a certain sense, minimized. Since H_0 and H_1 are composite hypotheses, the probability that a test will accept H_1 when H_0 is true depends upon which of the simple hypotheses that make up H_0 is true.

Neyman and Pearson [2] have proposed that a test, T_α for a hypothesis be termed *uniformly more powerful* than another test, T_β, if the probability for T_α of accepting the hypothesis if it is false, or the probability of rejecting it if it is true, does not exceed the corresponding probability for T_β no matter which of the simple hypotheses is actually true. Since there is no test which is uniformly more powerful than all other possible tests, it is usually required that a test be uniformly most powerful (UMP) among the members of some specified class of tests.

3. A symmetric test when the two universes have equal standard deviations. Let us consider, first, the hypothesis $M_y \geq M_x$ where the universes from which observations of varieties (X) and (Y), respectively, are drawn are normally distributed universes with equal standard deviations, σ, and means M_x and M_y respectively. Let us suppose a sample drawn of n random observations from the universe of variety (X) and a sample of n independent and random observations from the universe of (Y). The probability distribution of points in the sample space is given by

$$(1) \qquad p(x_1, \cdots, x_n; y_1, \cdots, y_n) = (2\pi\sigma^2)^{-n} e^{-\frac{1}{2\sigma^2}\left[\sum_i (x_i - M_x)^2 + \sum_i (y_i - M_y)^2\right]}.$$

In testing the hypothesis $M_y \geq M_x$, there is a certain symmetry between the alternatives (X) and (Y). If there is no *a priori* reason for choosing (X) rather than (Y), and if the sample point E_1: $(a_1, \cdots, a_n; b_1, \cdots, b_n)$ falls in the region of acceptance of H_0: then the point E_2: $(b_1, \cdots, b_n; a_1, \cdots, a_n)$ should fall in the region of acceptance of H_1. That is, if E_1 is taken as evidence that $M_y \geq M_x$; then E_2 can with equal plausibility be taken as evidence that $M_x \geq M_y$.

Any test such that E_1: $(a_1, \cdots, a_n; b_1, \cdots, b_n)$ lies in ω_0 whenever E_2: $(b_1, \cdots, b_n; a_1, \cdots, a_n)$ lies in ω_1 and vice versa, will be designated a symmetric test of the hypothesis $M_y \geq M_x$. Let Ω be the class of symmetric tests of H_0. If T_α is a member of Ω, and is uniformly more powerful than every other T_β which is a member of Ω, then T_α *is the uniformly most powerful symmetric test of* H_0.

The hypothesis $M_y \geq M_x$ possesses a UMP symmetric test. This may be shown as follows. From (1), the ratio can be calculated between the proba-

bility densities at the sample points $E: (x_1, \cdots, x_n; y_1, \cdots, y_n)$ and E': $(y_1, \cdots, y_n; x_1, \cdots, x_n)$. We get

$$\frac{p(E)}{p(E')} = \exp\left\{\frac{n}{\sigma^2}(\bar{x} - \bar{y})(M_x - M_y)\right\}, \tag{2}$$

where

$$\bar{x} = \frac{1}{n}\sum_i x_i, \qquad \bar{y} = \frac{1}{n}\sum_i y_i.$$

Now the condition $p(E) > p(E')$ is equivalent to $\frac{n}{\sigma^2}(\bar{x} - \bar{y})(M_x - M_y) > 0$. Hence $p(E) > p(E')$ whenever $(\bar{x} - \bar{y})$ has the same sign as $(M_x - M_y)$.

Now for any symmetric test, if E lies in ω_0, E' lies in ω_1, and vice versa. Suppose that, in fact, $M_y > M_x$. Consider a symmetric test, T_α whose region ω_0 contains a sub-region ω_{0U} (of measure greater than zero) such that $\bar{y} < \bar{x}$ for every point in that sub-region. Then for every point E' in ω_{0U}, $p(E') < p(E)$. Hence, a more powerful test, T_β could be constructed which would be identical with T_α, except that ω_{1U}, the sub-region symmetric to ω_{0U}, would be interchanged with ω_{0U} as a portion of the region of acceptance for H_0. Therefore, a test such that ω_0 contained all points for which $\bar{y} > \bar{x}$, and no others, would be a UMP symmetric test. This result is independent of the magnitude of $(M_x - M_y)$ provided only $M_y \geq M_x$. We conclude that $\bar{y} > \bar{x}$ *is a uniformly most powerful symmetric test for the hypothesis* $M_y > M_x$.

The probability of committing an error with the UMP symmetric test is a simple function of the difference $|M_y - M_x|$. The exact value can be found by integrating (1) over the whole region of the sample space for which $\bar{y} < \bar{x}$. There is no need to distinguish errors of the first and second kind, since an error of the first kind with T_0 is an error of the second kind with T_1, and vice versa. The probability of an error is one half when $M_x = M_y$, and in all other cases is less than one half.

4. Relation of UMP symmetric test and test which is UMP of tests absolutely equivalent to it. Neyman and Pearson [2] have shown the test $\bar{y} - \bar{x} > k$ to be UMP among the tests absolutely equivalent to it, for the hypothesis $M_y \geq M_x$. They have defined a class of tests as absolutely equivalent if, for each simple hypothesis in H_0, the probability of an error of the first kind is exactly the same for all the tests which are members of the class. If k be set equal to zero, $\bar{y} > \bar{x}$, and their test reduces to the UMP symmetric test. What is the relation between these two classes of tests?

If T_α be the UMP symmetric test, then it is clear from Section 2 that there is no other symmetric test, T_β, which is absolutely equivalent to T_α. Hence Ω, the class of symmetric tests, and Λ, the class of tests aboslutely equivalent to T_α, have only one member in common—the test T_α itself. Neyman and Pearson have shown T_α to be the UMP test of Λ, while the results of Section 4 show T_α to be the UMP test of Ω.

5. Justification for employing a symmetric test. In introducing Section 3, a heuristic argument was advanced for the use of a symmetric, rather than an asymmetric test for the hypothesis $M_y \geq M_x$. This argument will now be given a precise interpretation in terms of probabilities.

Assume, not a single experiment for testing the hypothesis $M_y \geq M_x$, but a series of similar experiments. Suppose a judgment to be formed independently on the basis of each experiment as to the correctness of the hypothesis. Is there any test which, if applied to the evidence in each case, will maximize the probability of a correct judgment in that experiment? Such a test can be shown to exist, providing one further assumption is made: that if any criterion be applied *prior* to the experiment to test the hypothesis $M_y \geq M_x$, the probability of a correct decision will be one half. That is, it must be assumed that there is no evidence which, prior to the experiment, will permit the variety with the greater yield to be selected with greater-than-chance frequency.

Consider now any asymmetric test for the hypothesis H_0—that is, any test which is not symmetric. The criterion $\bar{y} - \bar{x} > k$, where $k > 0$, is an example of such a test. Unlike a symmetric test, an asymmetric test may give a different result if applied as a test of the hypothesis H_0 than if applied as a test of the hypothesis H_1. For instance, a sample point such that $\bar{y} - \bar{x} = \epsilon$, where $k > \epsilon > 0$, would be considered a rejection of H_0 and acceptance of H_1 if the above test were applied to H_0; but would be considered a rejection of H_1 and an acceptance of H_0 if the test were applied to H_1. Hence, before an asymmetric test can be applied to a problem of dichotomous choice—a problem where H_0 *or* H_1 must be determinately selected—a decision must be reached as to whether the test is to be applied to H_0 or to H_1. This decision cannot be based upon the evidence of the sample to be tested—for in this case, the complete test, which would of course include this preliminary decision, would be symmetric by definition.

Let H_c be the correct hypothesis (H_0 or H_1, as the case may be) and let H_* be the hypothesis to which the asymmetric test is applied. Since by assumption there is no prior evidence for deciding whether H_c is H_0 or H_1, we may employ any random process for deciding whether H_* is to be identified with H_0 or H_1. If such a random selection is made, it follows that the probability that H_c and H_* are identical is one half.

We designate as the region of asymmetry of a test the region of points E_1: $(a_1, \cdots, a_n; b_1, \cdots, b_n)$ and E_2: $(b_1, \cdots, b_n; a_1, \cdots, a_n)$ of aggregate measure greater than zero such that E_1 and E_2 both fall in ω_0 or both fall in ω_1. Suppose ω_{0a} and ω_{0b} are a particular symmetrically disposed pair of subregions of the region of asymmetry, which fall in ω_0 of a test T_0. Suppose that, for every point, E_1, in ω_{0a}, $\bar{b} > \bar{a}$, and that ω_{0a} and ω_{0b} are of measure greater than zero. The sum of the probabilities that the sample point will fall in ω_{0a} or ω_{0b} is exactly the same whether H_c and H_* are the same hypothesis or are contradictory hypotheses. In the first case H_c will be accepted, in the second case H_c will be rejected. These two cases are of equal probability, hence there is a probability

of one half of accepting or rejecting H_c if the sample point falls in the region of asymmetry of T_0. But from equation (2) of Section 2 above, we see that if the subregions ω_{0a} and ω_{0b} had been in a region of symmetry, and if ω_{0a} had been in ω_0, the probability of accepting H_c would have been greater than the probability of rejecting H_c.

Hence, if it is determined by random selection to which of a pair of hypotheses an asymmetric test is going to be applied, the probability of a correct judgment with the asymmetric test will be less than if there were substituted for it the UMP symmetric test. It may be concluded that the UMP symmetric test is to be preferred unless there is prior evidence which permits a tentative selection of the correct hypothesis with greater-than-chance frequency.

6. Symmetric test when standard deviations of universes are unequal. Thus far, we have restricted ourselves to the case where $\sigma_x = \sigma_y$. Let us now relax this condition and see whether a UMP symmetric test for $M_y \geq M_x$ exists in this more general case.

We now have for the ratio of $p(E)$ to $p(E')$:

$$(3)\quad \frac{p(E)}{p(E')} = \exp\left\{-\frac{n}{2\sigma_x^2\sigma_y^2}\left[(\sigma_y^2 - \sigma_x^2)(\mu_x - \mu_y) - 2(\sigma_y^2 M_x - \sigma_x^2 M_y)(\bar{x} - \bar{y})\right]\right\},$$

where

$$\mu_x = \sum_i x_i^2/n, \qquad \mu_y = \sum_i y_i^2/n.$$

Even if σ_y and σ_x are known, which is not usually the case, there is no UMP symmetric test for the hypothesis $M_y \geq M_x$. From (3), the symmetric critical region which has the lowest probability of errors of the first kind for the hypothesis ($M_y = k_1$; $M_x = k_2$; $k_1 > k_2$) is the set of points E such that:

$$(4)\quad (\sigma_y^2 - \sigma_x^2)(\mu_x - \mu_y) - 2(\sigma_y^2 k_2 - \sigma_x^2 k_1)(\bar{x} - \bar{y}) > 0.$$

Since this region is not the same for all values of k_1 and k_2 such that $k_1 > k_2$, there is no UMP symmetric region for the composite hypothesis $M_y \geq M_x$. This result holds, *a fortiori* when σ_y and σ_x are not known.

If there is no UMP symmetric test for $M_y \geq M_x$ when $\sigma_y \neq \sigma_x$, we must be satisfied with a test which is UMP among some class of tests more restricted than the class of symmetric tests. Let us continue to restrict outselves to the case where there are an equal number of observations, in our sample, of (X) and of (Y). Let us pair the observations x_i, y_i, and consider the differences $u_i = x_i - y_i$. Is there a UMP test among the tests which are symmetric with respect to the u_i's for the hypothesis that $M_y - M_x = -U \geq 0$? By a symmetric test in this case we mean a test such that whenever the point $(u_1, \cdots, u_n)$ falls into region ω_0, the point $(-u_1, \cdots, -u_n)$ falls into region ω_1.

If x_i and y_i are distributed normally about M_x and M_y with standard deviations σ_x and σ_y respectively, then u_i will be normally distributed about $U =$

$M_x - M_y$ with standard deviation $\sigma_u = \sqrt{\sigma_x^2 + \sigma_y^2}$. The ratio of probabilities for the sample points $E_v: (u_1, \cdots, u_n)$ and $E_v': (-u_1, \cdots, -u_n)$ is given by:

$$\frac{p(E_v)}{p(E_v')} = \exp\left\{\frac{-2n}{\sigma_u^2}\,\bar{u}U\right\}, \tag{5}$$

where

$$\bar{u} = \frac{1}{n}\sum_i u_i .$$

Hence, $p(E_v) > p(E_v')$ whenever $\bar{u}$ has the same sign as U. Therefore, by the same process of reasoning as in Section 2, above, we may show that $\bar{u} \leq 0$ is a UMP test among tests symmetric in the sample space of the u's for the hypothesis $U \leq 0$.

It should be emphasized that Ω_{su}, the class of symmetric regions in the space of $E_v: (u_1 \cdots u_n)$, is far more restricted than Ω_s, the class of symmetric regions in the sample space of $E: (x_1 \cdots x_n ; y_1 \cdots y_n)$. In the latter class are included all regions such that:

(A) $E: (a_1, \cdots, a_n ; b_1, \cdots, b_n)$ falls in ω_0 whenever $E: (b_1, \cdots, b_n ; a_1, \cdots, a_n)$ falls in ω_1. Members of class Ω_{su} satisfy this condition together with the further condition:

(B) For all possible sets of n constants $k_1, \cdots, k_n$, $E: (x_1 + k_1, \cdots, x_n + k_n ; y_1 + k_1, \cdots, y_n + k_n)$ falls in ω_0 whenever $E: (x_1, \cdots, x_n ; y_1, \cdots, y_n)$ falls in ω_0. When $\sigma_y \neq \sigma_x$, a UMP test for $M_y \geq M_x$ with respect to the symmetric class Ω_{su} exists, but a UMP test with respect to the symmetric class Ω_s does not exist.

REFERENCES

[1] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Phil. Trans Roy. Soc.*, Series A, 702, Vol. 231 (1933), pp. 289–337.

[2] J. Neyman and E. S Pearson, "The testing of statistical hypotheses in relation to probabilities *A Priori*," *Proc. Camb. Phil. Soc.*, Vol. 29 (1933), pp. 492–510.

ON INDICES OF DISPERSION

By Paul G. Hoel

University of California, Los Angeles

1. Introduction. In biological sciences the index of dispersion for the binomial and Poisson distributions is very useful for testing homogeneity of certain types of data. For example, the dilution technique in making blood counts finds it useful. Recently there have been attempts to use it to determine allergies by observing the change in the blood count after allergic foods have been taken. Here the sample may consist of only a few readings, consequently it is important to know how accurate this index is when applied to small samples. After inspecting the application of the Poisson index to such counts, I was surprised to see the lack of agreement with theory. At first it appeared that the fault lay with the chi-square approximation which is used on this index, but later it was clear that the assumption of a basic Poisson distribution was at fault It now appears that statisticians will need to be careful about citing blood counts as examples of data following a Poisson distribution.

This paper is the result of investigating the accuracy of the chi-square approximation for the distribution of these indices. Previous work on this problem seems to have consisted in some sampling experiments [1] for small values of the parameters involved, and in some theoretical work [2] in which the sampling distribution is considered only for a fixed sample mean. Although sampling distributions ordinarily differ very little from the distributions obtained by assuming the mean of the sample fixed, for small degrees of freedom the difference may be appreciable and therefore requires investigation In this paper the accuracy of the chi-square approximation is investigated by finding expressions for the descriptive moments of the distribution which are correct to terms of order N^{-3}. These expressions are obtained by means of Fisher's semi-invariant technique.

2. Moments of the distribution. Employing Fisher's notation [3], let the binomial index of dispersion be denoted by z, then z may be written as:

$$z = \frac{\Sigma(x - \bar{x})^2}{\bar{x}\left(1 - \frac{\bar{x}}{n}\right)} = \frac{(N-1)k_2}{k_1\left(1 - \frac{k_1}{n}\right)} = \frac{N-1}{\kappa_1\left(1 - \frac{\kappa_1}{n}\right)} \frac{k_2}{\left(1 + \frac{k_1 - \kappa_1}{\kappa_1}\right)\left(1 - \frac{k_1 - \kappa_1}{n - \kappa_1}\right)}.$$

Letting $w = k_1 - \kappa_1$, $y = k_2$, $a = n - \kappa_1$, $b = \dfrac{N-1}{\kappa_1\left(1 - \frac{\kappa_1}{n}\right)}$, z may be ex-

panded as follows:

$$
\begin{aligned}
z &= \frac{by}{\left(1+\frac{w}{\kappa_1}\right)\left(1-\frac{w}{a}\right)} \\
&= by\left\{1-\frac{w}{\kappa_1}+\frac{w^2}{\kappa_1^2}-\cdots\right\}\left\{1+\frac{w}{a}+\frac{w^2}{a^2}+\cdots\right\} \\
&= by\left\{1+w\left(\frac{1}{a}-\frac{1}{\kappa_1}\right)+w^2\left(\frac{1}{a^2}-\frac{1}{a\kappa_1}+\frac{1}{\kappa_1^2}\right)+\cdots\right\} \\
&= b\{y+c_1wy+c_2w^2y+c_3w^3y+\cdots\},
\end{aligned}
$$

where the definition of c_i is obvious. As will be seen later, these expansions are valid for obtaining the expected values of powers of z; hence

$$
\begin{aligned}
E(z) &= b\ \{\mu_{01}+c_1\mu_{11}+c_2\mu_{21}+\cdots\} \\
E(z^2) &= b^2\{\mu_{02}+2c_1\mu_{12}+(2c_2+c_1^2)\mu_{22}+(2c_3+2c_2c_1)\mu_{32}+\cdots\} \\
E(z^3) &= b^3\{\mu_{03}+3c_1\mu_{13}+(3c_2+3c_1^2)\mu_{23}+(3c_3+6c_2c_1+c_1^3)\mu_{33}+\cdots\} \\
E(z^4) &= b^4\{\mu_{04}+4c_1\mu_{14}+(4c_2+6c_1^2)\mu_{24}+(4c_3+12c_2c_1+4c_1^3)\mu_{34}+\cdots\}.
\end{aligned} \tag{1}
$$

Since only the first four moments of z are to be found, it will be necessary to evaluate the μ_{ij} for $j = 1, 2, 3, 4$ and for $i = 0, 1, 2, \cdots$ as far as necessary to give the desired degree of accuracy.

First consider the relation between the moments μ_{ij} and the semi-invariants κ_{ij} which are defined in terms of the μ_{ij} by the following formal identity in t and τ.

$$
e^{\frac{\kappa_{10}t+\kappa_{01}\tau}{1!}+\frac{\kappa_{20}t^2+2\kappa_{11}t\tau+\kappa_{02}\tau^2}{2!}+\cdots} = 1+\frac{\mu_{10}t+\mu_{01}\tau}{1!}+\frac{\mu_{20}t^2+2\mu_{11}t\tau+\mu_{02}\tau^2}{2!}+\cdots.
$$

Differentiating both sides with respect to t and replacing the exponential factor by the right member gives an identity which is convenient for evaluating the μ_{i0}. Differentiating both sides with respect to τ and making the same replacement gives an identity which is convenient for evaluating the μ_{ij} for $j > 0$.

These identities express μ_{ij} as a sum of products of κ's and μ's, each such product being of total degree i and j in its subscripts. By repeated substitution, μ_{ij} can be expressed as a sum of products of κ's only. From Fisher's formulas

each such semi-invariant, κ_{rs}, can be expressed as a sum of products of semi-invariants of the basic distribution, each term of which sum is of order $N^{-(r+s-1)}$ in N. Hence it follows that the lowest order term, or at least one of the lowest order terms, in N in the expression for μ_{ij} will be a term with the maximum number of κ factors. Since the κ_{rs} of lowest degree in subscripts are κ_{10} and κ_{01}, the term with the maximum number of κ factors will be the term in $\kappa_{10}^{i}\kappa_{01}^{j}$. However, since $w = k_1 - \kappa_1$ has a zero mean value, $\mu_{10} = \kappa_{10} = 0$, consequently the lowest degree term involving the subscript $i > 0$ is κ_{20} or κ_{11}. As a result, the maximum number of κ factors will be found in the term containing $\kappa_{20}^{\frac{1}{2}i}\kappa_{01}^{j}$ for i even and $\kappa_{20}^{\frac{1}{2}(i-1)}\kappa_{01}^{j-1}\kappa_{11}$ for i odd. These terms are of order $N^{-\frac{1}{2}i}$ and $N^{-\frac{1}{2}(i+1)}$ respectively. Since it is desired to obtain accuracy of order N^{-3}, it therefore will suffice to evaluate μ_{ij} for $i \leq 6$.

The validity of the expansions used in arriving at (1) could now be shown by writing them as partial sums with remainder terms and then showing that the remainder terms are of higher order than N^{-3}.

Neglecting terms of higher order than N^{-3}, the above identities give the following expressions for μ_{ij} for $j = 0, 1, 2$ and $i = 0, 1, \cdots, 6$, with slightly longer expressions for $j = 3$ and 4.

$$\mu_{10} = 0 \qquad \mu_{01} = \kappa_{01}$$
$$\mu_{20} = \kappa_{20} \qquad \mu_{11} = \kappa_{11}$$
$$\mu_{30} = \kappa_{30} \qquad \mu_{21} = \kappa_{21} + \kappa_{01}\mu_{20}$$
$$\mu_{40} = \kappa_{40} + 3\kappa_{20}\mu_{20} \qquad \mu_{31} = \kappa_{31} + 3\kappa_{11}\mu_{20} + \kappa_{01}\mu_{30}$$
$$\mu_{50} = 6\kappa_{30}\mu_{20} + 4\kappa_{20}\mu_{30} \qquad \mu_{41} = 6\kappa_{21}\mu_{20} + 4\kappa_{11}\mu_{30} + \kappa_{01}\mu_{40}$$
$$\mu_{60} = 5\kappa_{20}\mu_{40} \qquad \mu_{51} = 5\kappa_{11}\mu_{40} + \kappa_{01}\mu_{50}$$
$$\mu_{61} = \kappa_{01}\mu_{60}$$

$$\mu_{02} = \kappa_{02} + \kappa_{01}\mu_{01}$$
$$\mu_{12} = \kappa_{12} + \kappa_{11}\mu_{01} + \kappa_{01}\mu_{11}$$
$$\mu_{22} = \kappa_{22} + \kappa_{21}\mu_{01} + \kappa_{02}\mu_{20} + 2\kappa_{11}\mu_{11} + \kappa_{01}\mu_{21}$$
$$\mu_{32} = \kappa_{31}\mu_{01} + 3\kappa_{12}\mu_{20} + 3\kappa_{21}\mu_{11} + \kappa_{02}\mu_{30} + 3\kappa_{11}\mu_{21} + \kappa_{01}\mu_{31}$$
$$\mu_{42} = 6\kappa_{21}\mu_{21} + \kappa_{02}\mu_{40} + 4\kappa_{11}\mu_{31} + \kappa_{01}\mu_{41}$$
$$\mu_{52} = 5\kappa_{11}\mu_{41} + \kappa_{01}\mu_{51}$$
$$\mu_{62} = \kappa_{01}\mu_{61}.$$

The next step is to apply Fisher's formulas expressing the κ_{rs} in terms of the semi-invariants of the basic variable distribution, which in this case is the binomial distribution. In Fisher's notation κ_{rs} would be written as $\kappa(1^r2^s)$, since the variables w and y are respectively k_1, measured from its expected value, and k_2. Applying such formulas, the following expressions for the μ_{i1} and μ_{i2} are obtained, with somewhat longer expressions for the μ_{i3} and μ_{i4}.

$$\mu_{01} = \kappa_2, \qquad \mu_{11} = \frac{\kappa_3}{N}, \qquad \mu_{21} = \frac{\kappa_4}{N^2} + \frac{\kappa_2^2}{N},$$

$$\mu_{31} = \frac{\kappa_5}{N^3} + \frac{4\kappa_3\kappa_2}{N^2}, \qquad \mu_{41} = \frac{7\kappa_4\kappa_2}{N^3} + \frac{4\kappa_3^2}{N^3} + \frac{3\kappa_2^3}{N^2},$$

$$\mu_{51} = \frac{25\kappa_3\kappa_2^2}{N^3}, \qquad \mu_{61} = \frac{15\kappa_2^4}{N^3}$$

$$\mu_{02} = \frac{\kappa_4}{N} + \kappa_2^2\left[\frac{2}{N-1} + 1\right]$$

$$\mu_{12} = \frac{\kappa_5}{N^2} + \frac{2\kappa_3\kappa_2}{N}\left[\frac{2}{N-1} + 1\right] \tag{2}$$

$$\mu_{22} = \frac{\kappa_6}{N^3} + \frac{\kappa_4\kappa_2}{N^2}\left[\frac{4}{N-1} + 3\right] + \frac{2\kappa_3^2}{N^2}\left[\frac{2}{N-1} + 1\right] + \frac{\kappa_2^3}{N}\left[\frac{2}{N-1} + 1\right]$$

$$\mu_{32} = \frac{5\kappa_5\kappa_2}{N^3} + \frac{7\kappa_4\kappa_3}{N^3} + \frac{7\kappa_3\kappa_2^2}{N^2}\left[\frac{2}{N-1} + 1\right]$$

$$\mu_{42} = \frac{16\kappa_4\kappa_2^2}{N^3} + \frac{20\kappa_3^2\kappa_2}{N^3} + \frac{3\kappa_2^4}{N^2}\left[\frac{2}{N-1} + 1\right]$$

$$\mu_{52} = \frac{40\kappa_3\kappa_2^3}{N^3}$$

$$\mu_{62} = \frac{15\kappa_2^5}{N^3}.$$

It is necessary to express these κ's in terms of the parameters of the binomial distribution. Here the κ's are defined by the following formal identity in θ,

$$e^{\kappa_1\theta + \kappa_2\frac{\theta^2}{2!} + \kappa_3\frac{\theta^3}{3!} + \cdots} = (q + pe^{\theta})^n.$$

Taking logarithms, expanding in powers of θ, and equating coefficients of powers of θ, the following expressions are obtained:

$$\kappa_1 = m$$

$$\kappa_2 = mq$$

$$\kappa_3 = mq(q - p)$$

$$\kappa_4 = mq(1 - 6pq)$$

$$\kappa_5 = mq(q - p)(1 - 12pq)$$

$$\kappa_6 = mq(1 - 30pq + 120p^2q^2)$$

$$\kappa_7 = mq(q - p)(1 - 60pq + 360p^2q^2)$$

$$\kappa_8 = mq(1 - 126pq + 1680p^2q^2 - 5040p^3q^3).$$

These values of the κ's are inserted in (2) to give the following expressions for the μ_{i1} and μ_{i2}, with considerably longer expressions for the μ_{i3} and μ_{i4}.

$$\mu_{01} = mq$$

$$\mu_{11} = mq\,(q - p)\,\frac{1}{N}$$

$$\mu_{21} = mq\left(\frac{1 - 6pq}{N^2} + \frac{mq}{N}\right)$$

$$\mu_{31} = mq(q - p)\left(\frac{1 - 12pq}{N^3} + \frac{4mq}{N^2}\right)$$

$$\mu_{41} = m^2 q^2\left(\frac{11 - 58pq}{N^3} + \frac{3mq}{N^2}\right)$$

$$\mu_{51} = m^3 q^3(q - p)\,\frac{25}{N^3}$$

$$\mu_{61} = m^4 q^4\,\frac{15}{N^3}$$

$$\mu_{02} = mq\left(\frac{1 - 6pq}{N} + \frac{2mq}{N - 1} + mq\right)$$

$$\mu_{12} = mq(q - p)\left(\frac{1 - 12pq}{N^2} + \frac{4mq}{N(N - 1)} + \frac{2mq}{N}\right)$$

$$\mu_{22} = mq\left(\frac{1 - 30pq + 120p^2q^2}{N^3} + \frac{8mq(1 - 5pq)}{N^2(N - 1)} + \frac{mq(5 - 26pq)}{N^2} + \frac{2m^2q^2}{N(N - 1)} + \frac{m^2q^2}{N}\right)$$

$$\mu_{32} = m^2 q^2(q - p)\left(\frac{12 - 102pq}{N^3} + \frac{14mq}{N^2(N - 1)} + \frac{7mq}{N^2}\right)$$

$$\mu_{42} = m^3 q^3\left(\frac{36 - 176pq}{N^3} + \frac{6mq}{N^2(N - 1)} + \frac{3mq}{N^2}\right)$$

$$\mu_{52} = m^4 q^4(q - p)\,\frac{40}{N^3}$$

$$\mu_{62} = m^5 q^5\,\frac{15}{N^3}.$$

It remains to express the coefficients of (1) in terms of these same parameters. From the definition of c_i, a, and κ_1, it follows that

$$c_i = \frac{\left(\frac{1}{a}\right)^{i+1} + (-1)^i\left(\frac{1}{\kappa_1}\right)^{i+1}}{\frac{1}{a} + \frac{1}{\kappa_1}} = \frac{p^{i+1} + (-1)^i q^{i+1}}{m^i q^i}.$$

If now the above values of the $\mu_{i,j}$ and c_i are inserted in the expressions (1), the following final formulas are obtained.

$$E(z) = (N-1)\left\{1 + \frac{p}{Nm} + \left(\frac{p}{Nm}\right)^2 + \left(\frac{p}{Nm}\right)^3 + \cdots\right\}$$

$$E(z^2) = (N-1)^2\left\{1 + \frac{2}{N-1} - \frac{2(1-6pq)}{(N-1)Nmq} + \frac{pq(2-11pq)}{(Nmq)^2} - \frac{2(1+2pq-25p^2q^2)}{(N-1)(Nmq)^2} + \frac{2pq(1+3pq-30p^2q^2)}{(Nmq)^3} + \cdots\right\}$$

$$E(z^3) = (N-1)^3\left\{1 + \frac{6}{N-1} - \frac{3pq}{Nmq} + \frac{8}{(N-1)^2} - \frac{6(1-3pq)}{(N-1)Nmq} + \frac{2pq(1-5pq)}{(Nmq)^2} + \frac{4(1-4pq)(N-2)}{(N-1)^2Nmq} - \frac{24(1-5pq)}{(N-1)^2Nmq} - \frac{6(1-11pq+40p^2q^2)}{(N-1)(Nmq)^2} + \frac{6pq(1-16pq+55p^2q^2)}{(Nmq)^2} + \frac{60pq(1-4pq)(N-2)}{(N-1)^2(Nmq)^2} + \cdots\right\} \tag{3}$$

$$E(z^4) = (N-1)^4\left\{1 + \frac{12}{N-1} - \frac{8pq}{Nmq} + \frac{44}{(N-1)^2} - \frac{12(1+2pq)}{(N-1)Nmq} - \frac{2pq(2-21pq)}{(Nmq)^2} + \frac{16(1-4pq)(N-2)}{(N-1)^2Nmq} + \frac{48}{(N-1)^3} - \frac{8(15-46pq)}{(N-1)^2Nmq} - \frac{12(3-44pq+138p^2q^2)}{(N-1)(Nmq)^2} + \frac{64pq(1-4pq)(N-2)}{(N-1)^2(Nmq)^2} + \frac{96(1-4pq)(N-2)}{(N-1)^3Nmq} + \frac{8(1-12pq+36p^2q^2)(4N^2-9N+6)}{(N-1)^3(Nmq)^2} + \frac{4pq(1-43pq+168p^2q^2)}{(Nmq)^3} + \cdots\right\}.$$

By considering the formation of terms, it can also be shown that the above expressions are correct to terms of order m^3, m^2, m^1, and m^0, respectively, in the parameter m. If m is large these expressions are considerably more accurate than the order N^{-3} would indicate since the lowest order terms neglected in these expressions are respectively N^4m^4, N^4m^3, N^4m^2, and N^4m.

3. Applications. To compare these moments with those of the chi-square distribution, consider the ratios of corresponding moments, both for the Poisson distribution and for the binomial distribution in the special case of $p = \frac{1}{2}$.

For the Poisson distribution, these ratios are

$$R_1 = 1$$

$$R_2 = 1 - \frac{1}{Nm} - \frac{1}{(Nm)^2}$$

$$R_3 = 1 + \frac{1}{2m} - \frac{4}{Nm}$$

$$R_4 = 1 + \frac{2}{N+3}\left\{\frac{3}{m} + \frac{1}{3m^2} - \frac{7}{Nm}\right\}.$$

For the binomial distribution with $p = \frac{1}{3}$, these ratios are

$$R_1 = 1 + \frac{1}{Nn} + \frac{1}{(Nn)^2} + \frac{1}{(Nn)^3}$$

$$R_2 = \left(1 - \frac{1}{n}\right)\left(1 + \frac{5}{2Nn} - \frac{7}{4N^2n^2}\right) - \frac{7}{4(Nn)^3}$$

$$R_3 = \left(1 - \frac{1}{n}\right)\left(1 - \frac{7}{4n}\right) + \frac{1}{Nn}\left(4 - \frac{13}{2n} + \frac{5}{2n^2}\right) + \frac{1}{N^2n^2}\left(1 - \frac{5}{n}\right) + \frac{5}{2(Nn)^3}$$

$$R_4 = 1 + \frac{N}{N+3}\left\{\frac{-2}{n} + \frac{1}{n^2} + \frac{1}{Nn}\left(-13 + \frac{37}{2n} - \frac{17}{2n^2}\right)\right.$$
$$\left. + \frac{1}{N^2n}\left(11 - \frac{67}{2n} + \frac{51}{2n^2}\right) + \frac{1}{N^3n^2}\left(\frac{31}{2} - \frac{51}{2n}\right) + \frac{17}{2N^4n^3}\right\}.$$

From these expressions the following table is constructed.

m	n	N	R_1	R_2	R_3	R_4
25	∞	3	1	.99	.97	1.01
25	75	3	1	1	.98	.97
5	∞	5	1	.96	.94	1.08
5	15	5	1.01	.96	.87	.84
2	∞	∞	1	1	1.25	1
2	∞	10	1	.95	1.05	1.19
2	∞	5	1	.89	.85	1.21
2	6	∞	1	.83	.59	.69
2	6	10	1.02	.87	.64	.64
2	6	5	1.03	.90	.69	.62
1	∞	25	1	96	1.34	1.22
1	∞	10	1	.89	1.10	1.39
1	∞	5	1	.76	.70	1.44
1	3	25	1.01	.69	.31	.41
1	3	10	1.03	.72	.35	.38
1	3	5	1.07	.77	.41	.36

For $m \geq 5$ these ratios are close to unity even for N as small as 5; hence it appears that the chi-square approximation is satisfactory as long as $m \geq 5$

For $m \leq 2$ most of these ratios differ considerably from unity, particularly for the binomial distribution. Since R_1 is practically constant, the reduction in R_2 here indicates that the chi-square approximation will contain too many extreme values For the Poisson distribution there is an increase in R_4 to compensate slightly for this decrease in R_2 so that the 5 percent points, for example, would not differ very much The use of the chi-square approximation would therefore tend to give slightly too few significant results when they exist. For the binomial distribution, however, there is a decrease in both R_2 and R_4, so that the distribution tends toward normality; consequently the chi-square approximation will contain far too many extreme values and the 5 percent point will be much too large. This situation becomes slightly worse with increasing N.

4. Conclusions. From a consideration of the approximations for the first four moments of the distribution of the index of dispersion, it appears that the chi-square approximation is highly satisfactory provided that $m \geq 5$. For smaller values of m, the approximation is still fairly accurate for the Poisson distribution but not for the binomial distribution. For decreasing small values of m there is an increasing tendency to claim compatibility between data and theory when it does not exist; hence the binomial index must be handled carefully in such situations These general conclusions are in agreement with the specialized results of Cochran and Sukhatme.

The semi-invariant technique for problems such as this is exceedingly laborious and is of questionable accuracy. The coefficients in Fisher's heavier formulas are so large that increased accuracy comes slowly with increased accuracy of order of terms. In addition, there are numerous typographical mistakes in Fisher's formulas, some of which are not easily detected. The formulas (3) may be used to investigate the accuracy of the chi-square approximation for situations not covered in the numerical table, but they are of questionable accuracy, when m is small, for N as small as 5.

REFERENCES

[1] P. V. Sukhatme, "On the distribution of chi-square in samples of the Poisson series," *Jour Roy. Stat. Soc.*, Vol. 101 (1938), pp. 75-79.

[2] W. G. Cochran, "The chi-square distribution for the binomial and Poisson series with small expectations," *Annals of Eugenics*, Vol. 7 (1936), pp. 207-17.

[3] R. A. Fisher, "Moments and product moments of sampling distributions," *Proc. London Math. Soc.*, Series 2, Vol. 30 (1930), pp. 199-238.

ON SERIAL NUMBERS

By E. J. Gumbel

New School for Social Research

In this paper we consider a continuous variate and unclassified observations. It is well known that there are two step functions, which we may trace for a given series of observations. We will show that the differences between the two ways of plotting play an important rôle for certain graphical methods used by engineers

To obtain one and only one series of observations we adjust the cumulative frequencies. The corrections thus introduced depend upon the theoretical distribution which is adequate for the observations Later we deal with the relation between serial numbers and grades. Finally we construct confidence bands for the comparison between theory and observations

1. Theory and observations. If we arrange n observations in order of increasing magnitude, and write each as often as it occurs, there will be a first, x_1, the smallest value, a second, x_2, an mth, x_m the penultimate, x_{n-1}, and the last, x_n, i.e, the greatest value. The index m is called the observed cumulative frequency, or simply the rank It is usual to draw the observations x_m along the abscissa, and the rank m along the ordinate. The step function starts with a vertical line from the value x_1 of the abscissa to the point with the coordinates 1, x_1, and in general consists of the horizontal lines from the point m, x_m to the point m, x_{m+1} and the vertical lines from the point m, x_{m+1} to the point $m + 1$, x_{m+1}. The step function ends with the point n, x_n. We call this graph the step function (m, x_m). However, another step function which is derived from the observations arranged in decreasing magnitude is equally legitimate This step function starts from the point with the coordinates 0, x_1, and in general consists of the horizontal lines from the point $m - 1$, x_m to $m - 1$, x_{m+1} and the vertical lines from the point $m - 1$, x_{m+1} to the point m, x_{m+1} and ends with the point $n - 1$, x_n We call it the step function $(m - 1, x_m)$. Let $F(x)$ be the probability of a value equal to or less than x. Then the continuous theoretical curve, the ogive, which we compare to the step functions is $nF(x)$, x. The question is whether we have to use the step function (m, x_m) or the step function $(m - 1, x_m)$.

The differences between the two ways of plotting are rarely mentioned in the statistical literature. If we plot instead of the rank m the reduced rank m/n, the differences between the two ways of plotting are of the order $1/n$. It is generally tacitly assumed that this difference may be neglected. This will not hold if n is small.

In the following we show two other ways of plotting the observations where the differences between the two observed curves play an important role. Sup-

pose that the probability $F(x)$ and the density of probability, $f(x)$, henceforth called the distribution are such that it is possible to introduce a reduced variate

$$z = \frac{x - a}{b}, \tag{1}$$

which has no dimension. In general, the constant a will be a certain mean, and the constant b a certain measure of dispersion. Furthermore, the constants may be linear functions of these characteristics. Neither the probability $G(z)$ of a value equal to or less than z

$$G(z) = F(x), \tag{2}$$

nor the reduced distribution

$$g(z) = bf(a + bz) \tag{3}$$

contain constants. The *equiprobability test* consists in the following procedure: We attribute to the mth observation x_m the relative frequency m/n, and determine from a probability table a value z, such that

$$G(z) = m/n. \tag{4}$$

The variate x is plotted on the ordinate, and the reduced variate z on the abscissa. Then the points x_m , z must be situated close to the straight line (1). To apply this comparison between theory and observations, we need not even calculate the constants. For the normal distribution the application of this test is greatly facilitated by the use of probability paper.

The difficulty is that we may as well choose the frequency

$$G(z) = (m - 1)/n, \tag{4'}$$

and determine the corresponding values of z. Therefore, we have two lines (1) instead of one. The difference between the two series will be large for the first and last few observations. For the first series the last observation cannot be plotted on probability paper; for the second series the first observation cannot be plotted.

The same difficulty exists for the "return period." If the observations of a continuous variate are made at regular intervals in time which are taken as units, we may as in [4] define the theoretical return periods $T(x)$ of a value equal to or greater than x as

$$T(x) = \frac{1}{1 - F(x)}. \tag{5}$$

The comparison of the theoretical with the observed return periods gives a test for the validity of a theory. However, there are two series of observations, namely, the exceedance intervals

$$(6) \qquad {}'T(x_m) = \frac{n}{n-m}; \qquad m = 1, 2 \cdots n-1$$

and the recurrence intervals

$$(7) \qquad {}''T(x_m) = \frac{n}{n-m+1}; \qquad m = 1, 2 \cdots n.$$

The two expressions (6) and (7) differ widely for the high ranks. The penultimate observation, for example, has an exceedance interval n, whereas the recurrence interval is only $n/2$. This contradiction and the difficulty arising for the equiprobability test show that the question of choosing the observed cumulative frequency of the mth observation has a practical significance.

The equiprobability test and the comparison between the observed and the theoretical return period may be combined on probability paper. The variate x is plotted on the ordinate, the reduced variate y on the abscissa. But instead of y we write the probability $F(x)$ and the return period $T(x)$. If the theory holds, the observations must be scattered around the straight line (1).

But all these methods presuppose that we know whether we have to attribute to x_m the rank m or the rank $m-1$. Sometimes a compromise has been proposed which consists in attributing to x_m neither m nor $m-1$, but the arithmetic mean of both, $m - \frac{1}{2}$. In other words, the index m is no longer considered to be an integer In such cases, we call m *the serial number*.

The corrected frequency $m - \frac{1}{2}$ may be accepted for the comparison between the step function and the probability curve. However, for the return period and for the equiprobability test this method leads to serious difficulties. The corrected return periods, which have been proposed by Hazen [7] and have been used by M. Kimball [8] are

$$(6) \qquad T(x_m) = \frac{n}{n-m+1/2}.$$

The last among n observations has a return period $2n$. This idea does not seem to be sound No statistical device can increase the number of observations beyond n.

2. The adjusted frequency of the mth observation. The use of m, $m-1$, or $m - \frac{1}{2}$ as frequency of the mth observations amounts to considering the mth value as being fixed. To obtain one and only one step function we consider x_m as a statistical variate. This will lead to the determination of the most probable serial number and of the corresponding probability as a function of m and n.

The mth observation is such that there are $m-1$ observations below it and $n-m$ observations above it. Consequently, the distribution $w_n(x, m)$ of the mth observation is

$$(9) \qquad w_n(x, m) = \binom{n}{m} m[F(x)]^{m-1}[1 - F(x)]^{n-m} f(x).$$

The variate x_m is simply called x as each value of x has a certain density of probability of being the mth. To distinguish between (x) and $w_n(x,m)$, the first distribution is referred to as the *initial* distribution. For some simple initial distributions it is possible to calculate exactly the mean and the standard error of the distribution (9). This has been done by Karl Pearson [10] for the normal, the uniform, the exponential, and other skew distributions. The results are very complicated, and do not allow any immediate practical applications

In the following we determine therefore instead of the mean the mode of the mth value. The most probable mth value for which we write $\tilde{x}_m$ is the solution of

$$\frac{d \log w_n(x,m)}{dx} = 0.$$

We obtain from (9)

$$\frac{m-1}{F(\tilde{x}_m)} f(\tilde{x}_m) - \frac{n-m}{1-F(\tilde{x}_m)} f(\tilde{x}_m) = -\frac{f'(\tilde{x}_m)}{f(\tilde{x}_m)}. \tag{10}$$

In this equation m is counted in order of increasing magnitude. If we choose the inverse order we obtain the same equation, if we replace the index m by $n - m + 1$. Therefore the following results are independent of the order of counting m.

Equation (10) gives the most probable value $\tilde{x}_m$ as a function of m and n. The function depends upon the distribution.

A rough, first trial solution of (10) may be obtained if we confine our interest to values where neither m nor $n - m$ is small in comparison to n, that is, values which are not extreme. Suppose m to be of the order $n/2$. For increasing numbers of observations, the expression on the left side of (10) become large compared to the right side provided the derivative remains finite, as is generally the case. If we neglect the right-hand member, $\tilde{x}_m$ is the solution of

$$F(\tilde{x}_m) = \frac{m-1}{n-1}. \tag{11}$$

This expression holds for the uniform distribution where $f'(x) = 0$.

The following exact solution is valid for any number of observations and any serial number Equation (10) will be used in two different ways: First, we suppose m to be known, we determine the probability $F(\tilde{x}_m)$ of the most probable mth value as a function of m and n, and attribute this probability to the mth observation x_m. By doing so, the probability of the most probable mth value becomes the *adjusted frequency* of the mth observation. This leads to one and only one series of observations, and settles our initial question. Later, in section 3, we suppose $F(\tilde{x}_m)$ to be known, and compute the corresponding most probable mth observation. This leads to an estimate of the grades (or partition values) from the serial numbers.

To obtain $F(\tilde{x}_m)$ from (10) we introduce an expression $\sigma^2(x_m)$ by stating

$$[\sigma^2(x_m)n] = F(\tilde{x}_m)[1 - F(x_m)]f^{-2}(\tilde{x}_m) . \tag{12}$$

The brackets are meant to indicate that the product on the left side does not depend upon n We shall show later that $\sigma^2(x_m)$ is under certain conditions the variance of the mth observation. For the present purpose however this significance is not required. Multiplication of (10) by (12) leads to

$$m - 1 + F(\tilde{x}_m) - nF(\tilde{x}_m) = -f'(\tilde{x}_m)[\sigma^2(x_m)n], \tag{13}$$

or

$$F(\tilde{x}_m) = \frac{m - 1}{n - 1} + \frac{f'(\tilde{x}_m)[\sigma^2(x_m)n]}{n - 1}. \tag{14}$$

The adjusted frequency in (14) is similar to (11). Another expression for the adjusted frequency, derived from (13), is

$$F(\tilde{x}_m) = \frac{m - \frac{1}{2}}{n} + \frac{1}{n}\,(F(\tilde{x}_m) - \tfrac{1}{2} + f'(\tilde{x}_m)[\sigma^2(x_m)n]). \tag{15}$$

The adjusted frequency is the compromise $\frac{m - \frac{1}{2}}{n}$ plus an expression

$$\frac{D}{n} = \frac{1}{n}\,(F(\tilde{x}_m) - \tfrac{1}{2} + f'(\tilde{x}_m)[\sigma^2(x_m)n]). \tag{16}$$

The correction, D, defined by (16) depends upon the initial distribution and has no dimension. In general, it will depend upon the constants which exist in the distribution. If the distribution $f(x)$ may be written in a reduced form (3), the correction[1]

$$D = G(z) - \tfrac{1}{2} + g'(z)[\sigma^2(z)n] \tag{17}$$

depends only upon the dimensionless reduced variate z. For a given initial distribution we choose numerical values for the probability $G(z) = F(\tilde{x}_m)$ calculate $g'(z)$ and

$$[\sigma^2(z)n] = \frac{G(z)(1 - G(z))}{g^2(z)}. \tag{18}$$

From (16) we compute a table for the corrections D as a function of the adjusted frequencies $F(\tilde{x}_m)$ and obtain for given n the serial number m as a function of the adjusted frequencies by

$$m = nF(\tilde{x}_m) + \tfrac{1}{2} - D. \tag{19}$$

These serial numbers will not be integers. The adjusted frequency $F(\tilde{x}_m)$ for the mth observation will then be obtained by linear interpolation.

[1] In previous articles [3, 6] we started from another interpretation of the corrected frequencies and obtained slightly different corrections.

The value and the sign of the correction D depends upon the distribution. For the asymmetrical exponential distribution, for example, the correction

$$D = -\tfrac{1}{2}, \tag{19'}$$

is independent of the variate. This means that we have to use exclusively the step function $(m - 1,\ x_m)$ as being the best way of plotting. The observed adjusted return periods are the recurrence intervals.

For a symmetrical reduced distribution we have

$$1 - G(-z) = G(z); \qquad g(-z) = g(z); \qquad g'(-z) = -g'(z). \tag{20}$$

Therefore, the reduced correction will be

$$D(-z) = -D(z). \tag{21}$$

For the two reduced values z and $-z$ of a symmetrical variate the corrections have the same size, but different signs.

A relation similar to (21) holds for two asymmetrical reduced distributions $g_1(z)$ and ${}_1g(z)$, which are symmetrical one to another in the sense

$$G_1(z) = 1 - {}_1G(-z), \qquad g_1(z) = {}_1g(-z); \qquad g_1'(z) = -{}_1g'(-z). \tag{22}$$

Then, the corrections are

$$D_1(-z) = -{}_1D(z). \tag{23}$$

For any initial distribution $f(x)$ we read from (19) the adjusted frequency

$$F(\tilde{x}_m) = \frac{m - \tfrac{1}{2} + D}{n}, \tag{24}$$

even for a small number of observations. The question whether to choose m/n or $(m - 1)/n$ as observed cumulative frequency is settled by (24). We obtain one observed step function, one series for the equiprobability test, and one series of observed return periods

$$T(\tilde{x}_m) = \frac{n}{n - m + \tfrac{1}{2} - D}, \tag{25}$$

which have to be compared to the theoretical continuous curves.

3. Estimates for the grades. In the following we use the fundamental formula (15) to determine interesting grades through the mth values.

We use the term *grade* for the value of a statistical variate which corresponds to a given cumulative probability $F(x)$ say, $F(x) = \frac{1}{4}; \frac{1}{2}; \frac{3}{4}$ for quartiles; $F(x) = \frac{1}{10}, \cdots \frac{9}{10}$ for deciles, and so on. For a given grade, the probability $F(x)$ the density of probability $f(x)$ and its derivative are known, and m is unknown. The value of m obtained from (15), henceforth called the most probable serial number $\tilde{m}$, is the solution of

$$\tilde{m} = nF(x) + 1 - F(x) - f'(x)F(x)(1 - F(x))f^{-2}(x). \tag{26}$$

The corresponding "observed" value $x_{\tilde{m}}$ is obtained by interpolation between two observed values x_{m-1} and x_m, such that

$$m - 1 < \tilde{m} < m.$$

For the median, x_0, the most probable serial number $\tilde{m}_0$ is

$$\tilde{m}_0 = \frac{n+1}{2} - \frac{f'(x_0)}{4f^2(x_0)}. \tag{27}$$

The median x_0 itself enters into (27). It has to be eliminated through the condition $F(x_0) = \frac{1}{2}$. For the exponential distribution for example we find

$$\tilde{m}_0 = \frac{n}{2} + 1. \tag{27'}$$

The most probable serial number of the median for a symmetrical distribution is

$$\tilde{m}_0 = \tfrac{1}{2}(n+1). \tag{28}$$

This is the usual estimate of the median for any distribution. The estimate obtained from (27) is smaller (larger) than the usual estimate if the median is smaller (larger) than the mode The difference between the two estimates is due to the fact, that (27) makes use of information about the theoretical distribution whereas this information (if available) is neglected by the usual method

For symmetrical distributions the most probable serial numbers $\tilde{m}_1$ and $\tilde{m}_2$ for two symmetrical grades defined by F_1 and $F_2 = 1 - F_1$ are according to (26) related by

$$\begin{aligned} \tilde{m}_1 &= nF_1 + 1 - (F_1 + f_1'F_1(1 - F_1)f_1^{-2}) \\ \tilde{m}_2 &= n(1 - F_1) + (F_1 + f_1'F_1(1 - F_1)f_1^{-2}). \end{aligned} \tag{29}$$

The members in brackets have the same size, but opposite signs. Another expression for $\tilde{m}_2$ is

$$\tilde{m}_2 = (n+1) - [nF_1 + 1 - F_1 - f_1'F_1(1 - F_1)f_1^{-2}]$$

so that, for symmetrical distributions

$$\tilde{m}_1 + \tilde{m}_2 = n + 1. \tag{30}$$

This is to be expected as the mth value counted upwards is the $(n - m + 1)$st value counted downwards.

For the two quartiles q_1 and q_2 the most probable serial numbers $\tilde{m}(q_1)$ and $\tilde{m}(q_2)$, obtained from (29) are

$$\tilde{m}(q_1) = \frac{n+3}{4} - \frac{3}{16}\frac{f'(q_1)}{f^2(q_1)}; \qquad \tilde{m}(q_2) = \frac{3n+1}{4} - \frac{3}{16}\frac{f'(q_2)}{f^2(q_2)}, \tag{31}$$

where q_1 and q_2 have to be eliminated by the use of

$$F(q_1) = \tfrac{1}{4};\ F(q_2) = \tfrac{3}{4}.$$

For the uniform, the normal and the exponential distribution we obtain the two quartiles from

$$\tilde{m}(q_1) = \frac{n+3}{4} \quad ; \qquad \tilde{m}(q_2) = \frac{3n+1}{4}$$

$$\tilde{m}(q_1) = \frac{n}{4} + .352; \qquad \tilde{m}(q_2) = \frac{3n}{4} + .648 \tag{31'}$$

$$\tilde{m}(q_1) = \frac{n}{4} + 1 \quad ; \qquad \tilde{m}(q_2) = \frac{3n}{4} + 1$$

respectively. The last result may also be found from (19') and (24). These estimates differ from the usual estimates by the reason given above.

We now apply the notion of a grade to certain characteristics which are *otherwise* defined. A certain characteristic, say, the mode $\tilde{x}$ or the mean $\bar{x}$ have for a given distribution the probabilities $F(\tilde{x})$ or $F(\bar{x})$ respectively. These probabilities may be used to define a grade. We determine the corresponding mth value from (26), and obtain an estimate of the mode or the mean, interpreted as grades by interpolation between the observed mth values For a symmetrical distribution these estimates of the mode and mean are identical with the estimates of the median. For an asymmetrical distribution, the most probable serial number $\tilde{m}(\tilde{x})$ of the mode becomes according to (26)

$$\tilde{m}(\tilde{x}) = (n-1)F(\tilde{x}) + 1. \tag{32}$$

Usually, the mode $\tilde{x}$ of a continuous variate is estimated by another procedure. The observations are arranged in certain cells. One of them has the largest relative frequency It will contain the mode. To find its position within the cell, an interpolation formula is applied which reproduces the content of this cell and of the two adjacent cells. By choosing different lengths for the cells and different origins for the classification, the mode can be shifted to the right or to the left. Formula (32) furnishes a determination of the mode from the observations according to the theory, such that the arrangement of the observations into different cells is not needed. Of course, this method can be applied only if we know the theoretical distribution $f(x)$. The problem how to estimate the mode is important for distributions where one of the constants may be interpreted as the mode or as a function of the mode [1, 4].

4. Standard errors of the estimates. The numerical work involved in the method (26) of estimating the grades is very small. To obtain the standard errors of these estimates we consider the asymptotic properties of the distribution (9). The following results hold therefore only for large numbers of observation. Besides we assume, that the serial number m is of the order $n/2$, i.e. not extreme. It has been shown [2] that under these conditions the distribution

of the mth value converges toward a normal distribution with a standard error $\sigma(x_m)$, where

$$\sigma(x_m)\sqrt{n} = \frac{1}{f(x)}\sqrt{F(x)(1-F(x))}. \tag{33}$$

Although this standard error does not contain m explicitly, it has a clear meaning for any value of x as we know from (26), which observation we have to attribute to the probability $F(x)$. The classical proof about the approximate normality of the distribution of the median in large samples is a special case of this convergence and the classical standard error of the median,

$$\sigma(x_0)\sqrt{n} = \frac{1}{2f(x_0)}, \tag{34}$$

is a special case of (33). The square root in (33) is maximum for $F(x) = \frac{1}{2}$. Therefore,

$$\sigma(x_m)\sqrt{n} \leqq \frac{1}{2f(x)}. \tag{35}$$

If the variate x may be reduced through the linear transformation (1) the standard error $\sigma(z)$ of the reduced variate, called reduced standard error

$$[\sigma(z)\sqrt{n}] = \frac{1}{g(z)}\sqrt{G(z)(1-G(z))}, \tag{36}$$

may be calculated as a function of z where z corresponds to x_m. To call attention to the fact that these numerical values do not depend upon n, they are written in brackets. The standard error of the estimate for x_m is, according to (2) and (3)

$$\sigma(x_m) = \frac{b}{\sqrt{n}}[\sigma(z)\sqrt{n}]. \tag{37}$$

Since the constant b is a measure of dispersion, the standard error of the estimate of the mth value is proportional to the standard deviation of the variate. The factors b and $1/\sqrt{n}$ show that the standard error of the mth value is of the same structure as the standard error of the arithmetic mean.

For symmetrical distributions the standard error (33) of the mth value is also a symmetrical function The standard errors of the estimate of the two quartiles, and generally of the estimates of two grades defined by F and $1 - F$, are then identical If the mode coincides with the median, the corresponding standard error of the mth value is a minimum. For a symmetrical U-shaped distribution, however, where the density of probability passes through a minimum at the center of symmetry, the median has the largest standard error among the mth values. An example for such a distribution has been given by Leavens [9]. As the distribution of the mth value converges towards a normal distribution, it is legitimate to attribute to the mode of the mth value the standard error (33)

Therefore, for a large number of observations (33) gives the standard error of our estimate of the grades. The standard errors of the estimates (31) of the quartiles are

$$\sigma(q_1)\sqrt{n} = \frac{\sqrt{3}}{4f(q_1)}; \qquad \sigma(q_2)\sqrt{n} = \frac{\sqrt{3}}{4f(q_2)}. \tag{38}$$

The arithmetic mean in its usual definition is not an mth value. Its standard error $\sigma(\bar{x})$, where

$$\sigma(\bar{x})\sqrt{n} = \sigma, \tag{39}$$

will, therefore, fall outside of the range of the standard errors of the mth values. (See graph 1.) If the distribution $f(x)$ is such that the standard deviation does not exist, it is legitimate to estimate the arithmetic mean as a grade, and calculate it from the corresponding most probable mth value by introducing $F(\bar{x})$, $f(\bar{x})$ and $f'(\bar{x})$ into (26). The standard error of the arithmetic mean interpreted as a grade is

$$\sigma(\bar{x})\sqrt{n} = \frac{1}{f(\bar{x})}\sqrt{F(\bar{x})(1 - F(\bar{x}))}. \tag{40}$$

If we use this estimate of the arithmetic mean for distributions where σ exists, the usual determination of the mean will be more (less) precise than its estimate as a grade if

$$\sigma f(\bar{x}) \lesseqgtr \sqrt{F(\bar{x})(1 - F(\bar{x}))}. \tag{41}$$

The standard error of the mode estimated as a grade is

$$\sigma(\tilde{x})\sqrt{n} = \frac{1}{f(\tilde{x})}\sqrt{F(\tilde{x})(1 - F(\tilde{x}))}. \tag{42}$$

As the standard error of any characteristic depends upon the way it is estimated from the observations, the standard errors of the mode or mean interpreted as grades differ from the usual standard errors.

5. The most precise grade. Equation (33) may be used to define a new grade which has interesting properties. The standard error (33) of the estimate of the mth value is a function of F. We ask whether it possesses a minimum (maximum). The corresponding value of the variate, $\hat{x}$, may be called *the most (least) precise mth value* or the most (least) precise grade. To obtain $\frac{d\sigma(x_m)}{dF}$ it is sufficient to calculate from (33)

$$\frac{nd \log \sigma^2(x_m)}{dx} = \frac{2n\sigma'(x_m)}{\sigma(x_m)}.$$

Therefore the most (least) precise grade is the solution of

$$\frac{f(x)}{F(x)} - \frac{f(x)}{1 - F(x)} - \frac{2f'(x)}{f(x)} = 0. \tag{43}$$

This expression does not vanish if either $F(x) = \frac{1}{2}$ or $f'(x) = 0$. It vanishes if both conditions hold simultaneously. For a symmetrical distribution passing through a mode (minimum), the mode (minimum), estimated as a grade, is the most (least) precise grade. Equation (43) may be written

$$f'(x)f^{-2}(x)F(x)(1 - F(x)) = \tfrac{1}{2} - F(x).$$

If we introduce this expression into (16), we obtain $D = 0$ and

$$F(\hat{x}) = \frac{m - \frac{1}{2}}{n}. \tag{44}$$

The most precise m*th value is such that the adjusted frequency is the arithmetic mean of the frequencies* m/n *and* (m − 1)/n.

The most precise mth value $\hat{x}$ cannot be calculated from the observations alone. It may be estimated in the same way as any grade by introducing the values $F(\hat{x})$, $f(\hat{x})$ and $f'(\hat{x})$ into equation (26).

To show the difference between the most precise grade and the mode we apply the procedure developed above to a skew distribution. The reduced distribution of the largest value $g(y)$ and the probability $G(y)$ are

$$g(y) = e^{-y}G(y); \qquad G(y) = e^{-e^{-y}}. \tag{45}$$

The relation (1) between the reduced variate for which we write y instead of z and the largest value x is

$$x = u + \frac{y}{\alpha}, \tag{46}$$

where $u = \tilde{x}$ is the mode and

$$\frac{1}{\alpha} = \frac{\sqrt{6}}{\pi}\sigma. \tag{47}$$

The most probable serial number $\tilde{m}(u)$ of the mode, obtained from (32) is

$$\tilde{m}(u) = \frac{n + e - 1}{e}. \tag{48}$$

This equation may be used for an estimate of the constant u.

The reduced variance $\sigma^2(y)$ obtained from (36) and (45) is

$$(\sigma^2(y)\sqrt{n}) = e^{2y}(e^{e^{-y}} - 1). \tag{49}$$

A table for the reduced standard error $\sigma(y)\sqrt{n}$ has been given in a previous publication [6]. The value $\sigma(y)\sqrt{n}$ is plotted in figure 1 for probabilities $G(y)$ from 0.01 to 0.95. The standard error has a minimum for a value of y located to the left of the mode $\tilde{y} = 0$. On the same graph are plotted the reduced standard errors for the normal distribution. As the normal reduced variate z differs from the reduced variate y, two different scales are used for the variates. The standard error of the estimate (48) of the mode interpreted as a grade, obtained by introducing $y = 0$ into (49) is

$$\sigma(u)\sqrt{n} = \frac{1}{\alpha}\sqrt{e-1} = 1.02205\sigma. \tag{49'}$$

The most precise grade is

$$\hat{x} = u + \frac{\hat{y}}{\alpha},$$

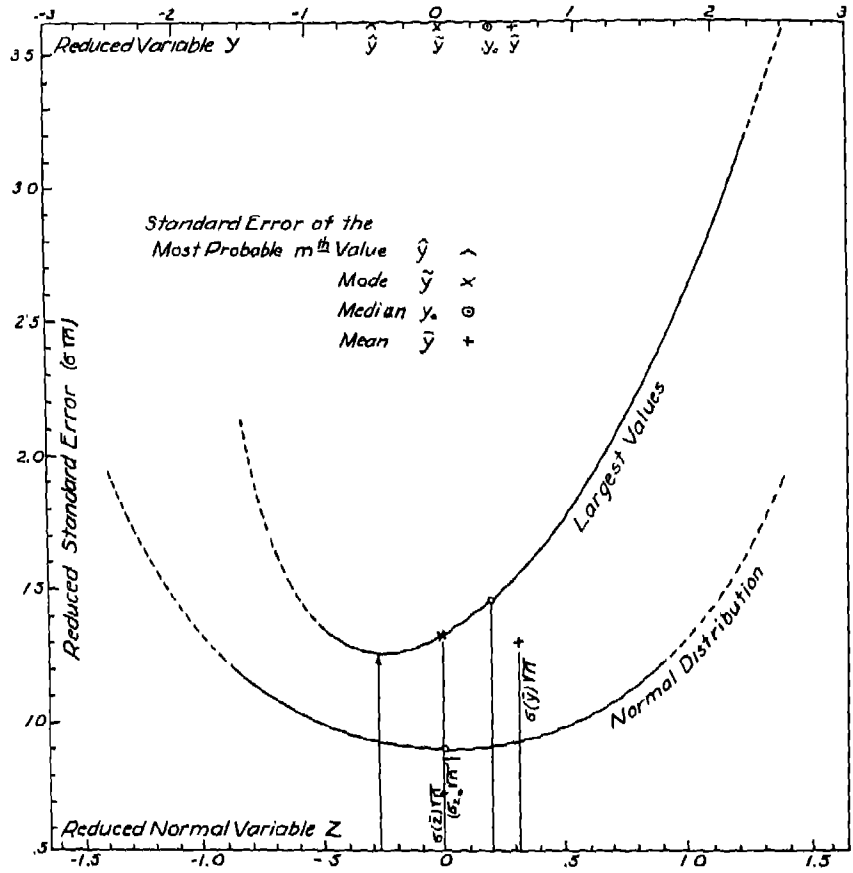


FIG. 1. The Reduced Standard Error of the mth Value

where $\hat{y}$ is the value of the reduced variate, for which the standard error (49) is minimum. We obtain from (49) and (45) the numerical values

$$\hat{y} = -.46601;\ G(\hat{y}) = .20319;\ \sigma(\hat{x})\sqrt{n} = .96887\sigma \tag{50}$$

The standard error of the most precise grade is 3 per cent smaller; the standard error of the mode, estimated as a grade, is 2 per cent larger than the standard error of the mean.

6. Confidence bands. The standard errors (33) of the grades may be used in a general way for the construction of *confidence bands* obtained from curves which control the fit between theory and observation. Consider first the observed stepfunction $(m - \frac{1}{2}, x_m)$ and the theoretical ogive $nF(x)$, x. The variate x is plotted along the abscissa, the cumulative frequency along the ordinate Now, for large n any theoretical value x, which is not extreme, may be interpreted as an mth value having a normal distribution and a standard error $\sigma(x_m)$ At each point of the graph of $nF(x)$, x which is not extreme, we construct a segment of length $2\sigma(x_m)$ parallel to the x axis, the midpoint of the segment being on the theoretical ogive. In other words, we add the standard error $\sigma(x_m)$ to, and subtract it from, any corresponding value x, and attribute $nF(x)$ to the beginning and end of these intervals. By this procedure we obtain two curves $nF(x)$, $x \mp \sigma(x_m)$ For each observation there exists a probability $P = .68268$ that it will be contained within the interval $x \mp \sigma(x_m)$.

If we apply another hypothesis to the same observations, or choose other values for the constants, we reach, of course, other control curves Of two competing hypotheses the one is to be preferred where the band contains a larger number of observations

The same method may be applied to the equiprobability test and to the comparison of the observed and theoretical return periods [6]. This procedure is legitimate for all values which are not extreme.

In the following, we construct the confidence bands for the normal distribution

$$g(z) = \frac{1}{\sqrt{\pi}} e^{-z^2}. \tag{51}$$

The variate x is related to the reduced variate z by (1), which, in this case, becomes

$$x = \bar{x} + \sigma\sqrt{2}z \tag{52}$$

The probability $G(z)$ is

$$G(z) = \tfrac{1}{2}(1 + \Phi(z)), \tag{53}$$

where $\Phi(z)$ stands for the Gaussian integral

$$\Phi(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2}\, dt. \tag{54}$$

Formulae (36) and (53) lead to the reduced standard error

$$\sigma(z)\sqrt{n} = \frac{1}{2g(z)} \sqrt{1 - \Phi^2(z)}, \tag{55}$$

given in the table, col 6. The standard errors $\sigma(x_m)$ of the mth values obtained from (37) (52) and (55) are

$$\sigma(x_m) = \frac{\sigma\sqrt{2}}{\sqrt{n}} [\sigma(z)\sqrt{n}]. \tag{56}$$

As a numerical example we choose the annual precipitations observed in 51 meteorological stations in Paris and its surroundings in the year 1938. We suppose that the differences between the 51 observations are only due to chance The stepfunction $m - \frac{1}{2}$, x_m is plotted in figure 2. To obtain the theoretical ogive we compute the constants in (52). They are

$$\bar{x} = 571.92; \sigma\sqrt{2} = 38.52. \tag{57}$$

The theoretical values x obtained from (52), the cumulative frequencies $nF(x)$ obtained from the table of the Gaussian integral [11] and the standard errors

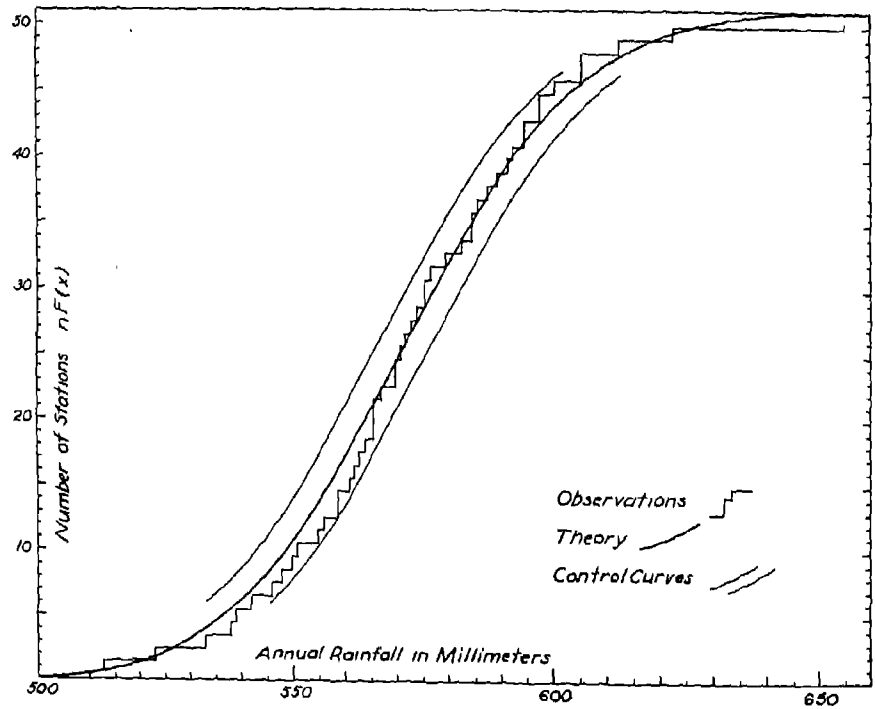


FIG. 2. The Confidence Band

$$\sigma(x_m) = 5.393\ [\sigma(z)\sqrt{\bar{n}}], \tag{58}$$

obtained from (56) are given in the columns 2 to 5 and 7 of Table I.

We trace in figure 2 the theoretical curve $nF(x)$, x and the confidence band obtained from col. 7. by the methods described above. All observations are contained within the control curves. We may accept the theory that the differences between the annual rainfalls observed in the 51 stations are only due to chance.

7. Conclusions. To test a statistical hypothesis for a continuous variate we use the ogive, the equiprobability method, based on (1), and the return periods

(5). The three tests may be combined on appropriate probability paper. As the rank of the mth observation x_m may be m or $m - 1$, we have two series of observations To obtain one and only one series we use for the ogive the serial number $m - \frac{1}{2}$ provided that the number of observations is large. Generally, we attribute to x_m an adjusted frequency, namely, the probability (15) of the most probable mth value. The adjusted frequency is obtained from the serial number $m - \frac{1}{2}$ and a correction, D, equation (17), which depends upon the distribution. The correction is important for the three tests, and small n, furthermore, for the equiprobability test and the return periods for the extreme observations and any number n.

The same correction D is used for estimating a grade through its relation (26) to the corresponding most probable serial number $\tilde{m}$. For distributions, where the second moment does not exist, we estimate the arithmetic mean from a

TABLE I

Normal Confidence Band and Theoretical Frequencies of the Rainfalls

Reduced Variate $\pm z$ (1)	Variate		Frequency		Reduced Standard Error $\sigma(z)\sqrt{n}$ (6)	Standard Error $\sigma(x_m)$ (7)
	x (2)	x (3)	$51\,F(x)$ (4)	$51\,F(x)$ (5)		
0	571.91	571.9	25.50	25.50	.886	4.8
.2	564.2	579.6	19.82	31.18	.899	4.9
.4	556.5	587.3	14.58	36 42	.940	5.1
.6	548.8	595 0	10.10	40.90	1.012	5.5
.8	541.0	602.7	6.58	44.42	1.127	6.1
1.0	533.4	610.4	4 01	46.99	1.297	7 0
1.2	525.7	618.1	2.29	48.71		
1.4	418.0	625.9	1.22	49 78		
1.6	510 3	633.6	60	50.40		
1.8	502.6	641.3	.28	50.72		

grade. For asymmetrical distributions we estimate the mode from a grade by (32) and (48).

In this case, we have to introduce a distinction between the mode and the most precise grade (43). The adjusted frequency and the estimates for grades may be used even for small numbers of observations

The standard error of these estimates is obtained, equation (33) from the limiting, normal, form of the distribution of the mth value, which holds, provided the serial number is not extreme. To control a hypothesis we construct confidence bands, which are obtained from the standard errors of the grades.

REFERENCES

[1] R A Fisher and L. H C Tippett, "Limiting forms of the frequency distribution of the largest or smallest member of a sample," *Proc. Camb. Phil Soc.*, Vol. 24, part 2 (1928), p. 180.

[2] E. J GUMBEL, "Les valeurs extrèmes des distribution statistiques," *Annales de l'Institut Henri Poincaré*, Vol 4 (1935), Paris, p. 115.
[3] E. J GUMBEL, "Les valeurs de position d'une variable aléatoire," *Comptes Rendus*, Vol. 208, (1939), Paris, p. 149.
[4] E. J. GUMBEL, "The return period of flood flows," *Annals of Math Stat* , Vol 12 (1942), p. 163.
[5] E. J. GUMBEL, "Simple tests for given hypotheses," *Biometrika*, Vol. 32 (1942), p. 317.
[6] E J GUMBEL, "Statistical control curves for flood discharge," *Trans. Am Geoph. Union* (1942), Washington, p. 489
[7] ALLEN HAZEN, *Flood Flows*, New York, John Wiley, 1930.
[8] B F. KIMBALL, "Limited type of primary probability distribution applied to annual flood flows," *Annals of Math. Stat.*, Vol 13 (1942), p. 318.
[9] DIXON H. LEAVENS, "Frequency distributions corresponding to time series," *Jour Amer Stat. Assoc* , Vol 26 (1931), p 407
[10] KARL AND MARGARET V. PEARSON, "On the mean character and variance of a ranked individual, and on the mean and variance of the interval between ranked individuals," *Biometrika*, Vol 23, part 3, 4 (1931), p. 364, Vol 24, part 1, 2 (1932), p 203
[11] *Tables of Probability Functions*, Federal Works Agency W.P A of New York City, 1941.

FITTING GENERAL GRAM-CHARLIER SERIES

By Paul A. Samuelson

Massachusetts Institute of Technology

1. Introduction. Since the last part of the nineteenth century at least, it has been common to represent a probability distribution by means of a linear sum of terms consisting of a parent function and its successive derivatives. Usually the parent function is the Type A or normal curve, as discussed by Gram [1], Bruns [2], Charlier [3], and numerous others. In addition there have been generalizations in various directions: for example, the Type B expansion in terms of the Poisson parent function and its successive finite differences.

Unlike these two types, which have a definite probability interpretation, another generalization involves the use of other parent functions and their derivatives (or differences) to give an approximate representation of a given frequency curve. With this process is associated the names of Charlier, Carver [4], Roa [5], and many others. Two general methods by which the equating of moments of the fitted curve and the given distribution yield the appropriate coefficients have been given by Charlier and Carver respectively. An account of the latter's technique is more accessible to the average English speaking statistician

It is the purpose of the present discussion to indicate how the Charlier method may be simplified, and can be used to replace the Carver method. In doing so, I am following up the oral suggestion made some years ago by Professor E. B. Wilson of Harvard, that repeated integration by parts will yield the requisite coefficients very simply. At the same time certain methods implicit in the work of Dr A C. Aitken [6] show how the use of a moment generating function can often lighten the algebraic analysis. There will also be a brief indication of analogous results for general finite difference parent families; and attention will be called to a troublesome historical blunder which has permeated the statistical literature.

2. Alternative methods. Avoiding the overburdened 'expression generating function, I shall consider parent functions, called $f(x)$, with the restrictive properties:

a) Moments of all order of $f(x)$ exist.

b) Derivatives of any required order exist with appropriate continuity.

c) There exist high order contact at the extremities of the distribution as defined below.

Mathematically,

a) $\int_{-\infty}^{\infty} x^k f(x)dx$ is finite for all positive integral values of k

and

c) $\lim_{x\to\pm\infty} x^j f^k(x) = 0$ for all positive integral values of j and k.

These conditions suffice for many statistically interesting cases, but where desirable they can be lightened. Thus, derivatives may only be defined "almost everywhere," and there may be finite instead of infinite limits to the distribution, etc.

Given an arbitrary frequency curve $F(x)$, we shall suppose it to be formally expanded in the series

$$(1) \qquad F(x) \sim a_0 f(x) + a_1 f'(x) + a_2 f''(x) + \cdots + a_n f^n(x) + \cdots .$$

For convenience in what follows, we shall assume that all distributions are given in terms of relative frequency so that the area under both f and F is equal to unity, so that a_0 may be taken as unity. The suppressed absolute frequency can clearly be restored at any time by multiplication of both sides with the appropriate constant. Also for algebraic convenience, many writers consider the slightly modified form of the expansion

$$F(x) \sim A_0 f(x) - \frac{A_1}{1!} f'(x) + \frac{A_2}{2!} f''(x) + \cdots \frac{(-1)^n A_n}{n!} f^n(x) + \cdots .$$

It is assumed without discussion that the first n coefficients in such a series are to be determined by equating the first n moments of each side.

I shall prove the two following identities:

$$(2) \qquad (-1)^n a_n = L_n(F) - \sum_0^{n-1} L_{n-i}(f)(-1)^i a_i ,$$

where

$$L_i(f^i) = \frac{\int_{-\infty}^{\infty} x^i f^i(x) dx}{i!} .$$

Alternatively

$$(3) \qquad A_n = \sum_{i=0}^{n} \binom{n}{i} \frac{d^i}{d\alpha^i} \left(\frac{1}{\int_{-\infty}^{\infty} f e^{\alpha x} \, dx} \right)_{\alpha=0} \int_{-\infty}^{\infty} x^{n-i} F(x) \, dx.$$

The first of these which I owe to Prof. Wilson is implicit in Charlier's work. The second which may fairly be attributed to Aitken may reduce the actual work in many special cases met in practice.

Both of these methods are closely related to the Charlier device of finding polynomials $S_n(x)$ with the bi-orthogonal property

$$\int_{-\infty}^{\infty} S_n(x) f^i(x) \, dx = 0, \qquad i \neq n.$$

The subscript indicates the degree of the polynomial. By means of n of the above relationships, the polynomials can be determined except for a factor of

proportionality. By formal integration of both sides of our expansion we have the Charlier identity

$$a_n = \int_{-\infty}^{\infty} S_n(x)F(x)\,dx/\text{factor of proportionality}.$$

From a theoretical standpoint, this method leaves little to be desired; but in practice the algebraic work increases rapidly with the number of terms to be included in the series.

In the Carver method, the new parent function in question, as well as the function to be approximated, are both expanded in terms of the normal curve, thus almost doubling the numerical calculations After some differentiation, the members of the Type A family are eliminated yielding in the process the required coefficients in terms of the new parent family. We shall see below how this method may be related to the three above.

3. Useful relationship. First, two simple identities may be presented:

$$\begin{aligned} L_j(f^{(i)}) &= (-1)^i L_{j-i}(f), \quad j \geqq i \\ &= 0 \qquad\qquad , \quad j < i. \end{aligned}$$

Given the above assumptions of high contact, this follows immediately from repeated integration by parts.

Remembering that the reduced moments defined just above are the coefficients of the powers of α in the series expansion of the moment generating function

$$M(\alpha; f^{(i)}) = \int_{-\infty}^{\infty} e^{\alpha x} f^{(i)}(x)\,dx = L_0(f^{(i)}) + L_1(f^{(i)})\alpha + L_2(f^{(i)})\alpha^2 + \cdots$$

we have the useful Aitken identity

$$M(\alpha; f^{(i)}) = (-1)^i M(\alpha; f)\alpha^i. \tag{4}$$

This, too, is the immediate consequence of repeated integration by parts.

4. Derivation of first method. Formally multiplying each side of (1) by $x^n/n!$ and integrating, we have the formal identity

$$L_n(F) = a_0 L_n(f) - a_1 L_{n-1}(f) + \cdots + (-1)^n a_n L_0(f)\,.$$

This is a "triangular" system of linear equations in the unknown a's. It may be written in matrix terms

$$\begin{bmatrix} L_0(F) \\ L_1(F) \\ \cdot \\ \cdot \\ \cdot \end{bmatrix} = \begin{bmatrix} L_0(f) & 0 & 0 & \cdots \\ L_1(f) & L_0(f) & 0 & \cdots \\ L_2(f) & L_1(f) & L_0(f) & \cdots \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \end{bmatrix} \begin{bmatrix} a_0 \\ -a_1 \\ a_2 \\ -a_3 \\ \cdot \\ \cdot \end{bmatrix}.$$

The triangular matrix has the very special property that all of its elements are known as soon as the first column is given. For this reason, as we shall see, it is essentially equivalent to a simple sequence of numbers. This we shall call the *sequence property*. Because of this special form, the above system by simple rearrangement may be written in the modified form

$$\begin{bmatrix} L_0(F) & 0 & \cdots \\ L_1(F) & L_0(F) & \cdots \\ \cdot & & \\ \cdot & & \\ \cdot & \cdot & \\ \cdot & \cdot & \end{bmatrix} = \begin{bmatrix} L_0(f) & 0 & \cdots \\ L_1(f) & L_0(f) & \cdots \\ \cdot & \cdot & \\ \cdot & \cdot & \\ \cdot & \cdot & \\ \cdot & \cdot & \end{bmatrix} \begin{bmatrix} a_0 & 0 & 0 \cdots \\ -a_1 & a_0 & 0 \cdots \\ a_2 & -a_1 & a_0 \cdots \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \end{bmatrix}$$

By appropriate definition of symbolism, this may be written in the simple matrix form:

$$L(F) = L(f)\, a(F, f),$$

since multiplication of two triangular, "sequence" matrices is commutative.

It is usually simplest to invert this triangular solution directly as in (2). But if necessary, we may express our answer in the equivalent form

$$a(F, f) = L(F)\, L(f)^{-1}, \tag{5}$$

where the inverse to any special triangular matrix, also possesses the sequence property.

If g is a second parent function with the properties of Section 2, we have the relationship

$$a(F, g) = a(F, f)\, a(f, g)$$

which follows directly from (5). This may be generalized to

$$a(f_1, f_2)\, a(f_2, f_3) \cdots a(f_{n-1}, f_n) = a(f_1, f_n)\,.$$

If F itself is a parent function, we have

$$a(F, f)\, a(f, F) = a(F, F) = I$$

or

$$a(f, F) = a(F, f)^{-1}\,.$$

6. Relation to old methods. In terms of our notation, the Carver method seems to reduce to computing $a(F, f)$ by the relationship

$$a(F, f) = a(F, \phi)\, a(f, \phi)^{-1}$$

where ϕ is the Type A parent function. It involves a doubling of the work of coefficient determination. However, if only a few terms in the expansion are retained, this is of negligible importance.

The Charlier polynomials are clearly summed rows of the matrix product

$$L(f)^{-1} \cdot \begin{bmatrix} 1/1! & 0 & 0 & \cdots \\ 0 & x/1! & 0 & \cdots \\ 0 & 0 & x^2/2! & \cdots \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \end{bmatrix}.$$

To know the first n of these polynomials, it is not necessary to derive $n(n+1)/2$ different coefficients. Because of the sequence property, it is only necessary to derive n elements of the first column of $L(f)^{-1}$ These can be expressed in terms of the reduced moments of f, as did Charlier; but the relationships are non-linear and algebraically become tedious for high n. They are better computed from sequence relationships.

The above discussion suggests that the bi-orthogonal relationship between a parent family and suitable polynomials has no deep significance. In particular, there is no essential relationship to least squares as in orthogonal expansions. It does, however, share one important property with orthogonal functions—determination of later coefficients does not affect the earlier ones. But this is a property of all triangular reductions, orthogonal expansions being only special cases of these.

6. Sequence properties. Ordinarily to derive the inverse of an n^2 matrix, n^2 equations must be solved For our triangular matrices, we need only solve n equations for one column. To each triangular matrix $L(f)$ there corresponds a sequence $\{L_k(f)\}$, which is in fact the first column of the former. Similarly to $L(f)^{-1}$, there corresponds $\{\bar{L}_k(f)\}$; the elements of the latter are defined by the n equations

$$\begin{aligned} &L_0(f)\bar{L}_0(f) = 1 \\ &L_0(f)\bar{L}_1(f) + L_1(f)\bar{L}_0(f) = 0 \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &\sum_0^n L_k(f)\bar{L}_{n-k}(f) = 0 \end{aligned}$$

But these are precisely the equations involved in the formal inversion of any linear operator system of the form

$$\sum_0^\infty c_k h^k y = z \tag{6}$$

where h is an operator which commutes with a constant, and for which $h^0 = 1$. z is a known function and y unknown. Thus h may be such operators as

$$x,\ d/dx,\ xd/dx,\ E,\ \Delta.$$

A particular solution of (6) is given by the formal expansion

$$y = \sum_0^\infty \bar{c}_i h^i z$$

where the $\bar{c}$'s bear the same relationship to the c's as do the $\bar{L}$'s to the L's.

Such "reciprocal" sequences appear in many branches of applied mathematics. In particular, they arise in the inversion of a power series. If formally,

$$W(\alpha) = \sum_0^\infty S_k \alpha^k$$

then

$$\frac{1}{W(\alpha)} = \sum_0^\infty \bar{S}_k \alpha^k.$$

Thus, to any triangular matrix with the sequence property, we can formally associate a function $W(\alpha)$ as well as a sequence of numbers. The calculus of multiplication of our triangular matrices clearly "corresponds" to the calculus of multiplication of functions, i.e if the triangular matrices T_1, T_2, $\cdot\ \cdot\ T_n$ and $W_1(\alpha)$, $W_2(\alpha)$, $\cdot\ \cdot\ W_n(\alpha)$ correspond, and $T_n = T_1 \cdot T_2 \cdots T_{n-1}$; then

$$W_n(\alpha) = W_1(\alpha)W_2(\alpha) \cdots W_{n-1}(\alpha).$$

Also, $1/W_1(\alpha)$ corresponds to T_1^{-1}.

7. Moment generating functions. If only for the above reasons and no others, we should be tempted to consider the function formally defined by

$$\sum_0^\infty L_k(f)\alpha^k.$$

But this is precisely the expression for the familiar moment generating function, m. g. f.

$$M(\alpha; f) = \int_{-\infty}^{\infty} e^{\alpha x} f(x)\, dx = \sum_0^\infty L_k(f)\alpha^k.$$

In this way, the method of triangular matrices joins the method used by Aitken for the Type A family. If

$$F(x) \sim \sum_0^\infty a_i f^i(x),$$

and we formally equate moment generating functions of each side, we get

$$M(\alpha; F) = M(\alpha; f) \sum_0^\infty (-1)^i a_i \alpha^i, \tag{7}$$

by means of the Aitken identity (4). Thus $(-1)^i a_i$ equals the coefficient of α^i in the formal expansion of

$$\frac{M(\alpha, F)}{M(\alpha; f)} = M(\alpha; F)M(\alpha; f)^{-1}.$$

Our relationship (2) follows immediately from (7); and by Taylor's expansion in α of $M(\alpha; f)^{-1}$, the identity (3) is quickly realized

For many problems, the reciprocal of the m. g. f. of $f(x)$ is itself a simple function; to that our triangular equations may be inverted without solving linear equations. Thus where $F(x) = f(x + b)$, we immediately verify Taylor's expansion by use of familiar properties of the m. g. f. under shift of origin.

8. Finite difference expansions. Corresponding to integration by parts, we have the formula

$$\sum_{-\infty}^{\infty} W_i \nabla^k V_i = (-1) \sum_{-\infty}^{\infty} \Delta W_i \nabla^{k-1} V_i = (-1)^2 \sum_{-\infty}^{\infty} \Delta^2 W_i \nabla^{k-2} V_i, \quad \text{etc.},$$

provided "high contact" properties are assumed. ∇ and Δ are receding and advancing differences respectively. Recalling the familiar property of "reduced factorial" polynomials, ${}^k x$, we have

$$\sum_{-\infty}^{\infty} {}^j x \nabla^k f(x) = (-1)^k \sum_{-\infty}^{\infty} {}^{j-k} x f(x) \qquad j \geqq k$$

$$= 0 \qquad j < k,$$

or

$$Q_j(\nabla^k f) = (-1)^k Q_{j-k}(f) \qquad j \geqq k$$

$$= 0 \qquad j < k,$$

where

$$Q_j(g) = \sum_{-\infty}^{\infty} \frac{x(x-1)(x-2)\cdots(x-j+1)}{j!}\, g(x).$$

In the expansion

$$F(x) \sim a_0 f(x) + a_1 \nabla f(x) + a_2 \nabla^2 f(x) + \cdots,$$

the a's obey laws identical to (2) and (3) where reduced factorial moments are substituted for the reduced L moments, and the f. m. g. f.

$$\sum_{-\infty}^{\infty} f(x)(1 + \alpha)^x,$$

for the ordinary m. g. f.

9. Convergence. All of the above relationships are purely formal, without regard to convergence. The last is a difficult subject, and little discussed in the statistical literature, since applications of G-C series have been almost entirely concerned with empirical frequency curve fitting in which mathematical con-

vergence does not enter. Actually in the scanty treatments of the subject there has arisen a confusion between the Type A *G-C* expansion, which equates moments, and the expansion of a function in orthogonal Hermite functions. These are not unrelated, but nevertheless they are distinct. This is well recognized in the purely mathematical literature, but hardly at all in the literature of statistics and physics.

The series differ by an irremovable factor of 2. If the Type A functions are written as

$$H_i(x)e^{-x^2},$$

then the Hermite functions will take the form

$$H_i(x)e^{-\frac{1}{2}x^2},$$

where the H's are Hermite polynomials suitably normalized. Unfortunately the *G-C* series often diverges when the H series converges Thus, the statistically interesting Cauchy distribution can be expanded in an H series; but since it possesses no finite higher moments, the *G-C* series cannot even be defined.

It is not hard to show that the *G-C* expansion of F in terms of a Type A function $f(x)$, is equivalent to an H expansion of $Ff^{-\frac{1}{2}}$ in terms of the H family $f^{\frac{1}{2}}$. It is sufficient for convergence in the mean of the last expansion that $Ff^{-\frac{1}{2}}$ be of integrable square or belong to L^2. This means that the *G-C* type A expansion will be valid if $Ff^{-\frac{1}{2}}$ is well behaved, not simply if F is well behaved. For F a histogram as is often the case in practise, no difficulties of convergence arise, although rapid convergence may be another matter Nevertheless, many well behaved F's will not pass the more strict test. The reader is referred to the last five titles in the bibliography for mathematical discussions of this problem.

The above discussion holds only for the Type A expansion. There remains the very difficult problem of convergence conditions in the more general case. No immediate generalization suggests itself, except the application of the results of the "moment problem" However, this must be handled with delicacy, since the partial sums of the series may actually become negative over some range.

BIBLIOGRAPHY

References

[1] J P. Gram, "Ueber die Entwicklung reeler Functionen in Reihen mittelst der Methode der Kleinsten Quadrate," *Journal für die Reine und Angewandte Mathematik*, Vol. 94 (1882), pp. 41-73.

[2] H Bruns, *Wahrscheinlichkeitsrechnung und Kollektivmasslehre*, B G Teubner, Leipzig, 1906

[3] C V. L. Charlier, "Über die Darstellung willkürlicher Funktionen," *Arkiv för Matematik, Astronomi och Fysik (utgivet af k. svenska vetenskapsakademien)*, Vol 2 (1905), 35 pp.

[4] H C Carver, Chapter VII, *Handbook of Mathematical Statistics*, edited by H L. Rietz, Cambridge, Massachusetts, 1924.

[5] Emeterio Roa, "A number of new generating functions with application to statistics," 1923

[6] A C AITKEN, *Statistical Mathematics*, London, 1939.
[7] V ROMANOVSKY, "Generalisation of some Types of the Frequency Curves of Professor Pearson," *Biometrika*, Vol 16 (1924), pp 106-116
[8] DUNHAM JACKSON, *Fourier Series and Orthogonal Polynomials*, (Carus Mathematical Monograph No. 6), The Mathematical Association of America, Oberlin, Ohio, 1941.
[9] E. H. HILDEBRANDT, "Systems of polynomials connected with the Charlier expansions and the Pearson differential and difference equations," *Annals of Math Stat*, Vol II, 1931, pp 379-439.

Further Literature

1 W MYLLER-LEBEDEFF, "Die Theorie der Integralgleichungen in Anwendung auf einige Reihenentwicklungen," *Math Annalen*, Vol 64 (1907), pp 388-416.
2 E HILLE, "A class of reciprocal functions," *Annals of Math*, Vol. 27 (1926), pp. 427-464. (Contains selected bibliography)
3 M H. STONE, "Developments in Hermite polynomials," *Annals of Math.*, Vol 29 (1927), pp 1-13.
4 W E. MILNE, "On the degree of convergence of the Gram-Charlier series," *Trans Amer Math Soc*, Vol 31 (1929), pp. 422-444.
5 *Bibliography on Orthogonal Polynomials*, National Research Council of the National Academy of Sciences, Washington, D C., 1940.

A METHOD OF TESTING THE HYPOTHESIS THAT TWO SAMPLES ARE FROM THE SAME POPULATION

By Harold C. Mathisen

Princeton University

1. Introduction. There are many cases in testing whether two samples are from the same population in which no assumption about the distribution function of the population can be made except that it is continuous. A. Wald and J. Wolfowitz, [1], have developed a method of testing the hypothesis that two samples come from the same population based on certain kinds of runs of the elements from each sample in the combined ordered sample. W J. Dixon, [2], has introduced a criterion for testing the same hypothesis based on the number of elements of the second sample falling between each successive pair of ordered values in the first sample.

The problem considered here is that of devising a simple method of testing the hypothesis that two samples come from the same population, based on medians and quartiles, given only that the distribution function of the population is continuous. The simplest method may be described briefly as follows. We observe the number of elements, m_1, in the second sample whose values are lower than the median of the first sample. Since the distribution of m_1 is independent of the population distribution, we are able to compute significance points from the distribution of m_1. These points may then be used for testing the hypothesis at a given significance level. This will be referred to as the case of two intervals.

This method may be easily extended to the case of any number of intervals. In this note we shall consider the extension to four intervals by using the median and the two quartiles of the first sample to establish four intervals into which the elements of the second sample may fall. Then, if the second sample is of size $4m$, it will be shown that, under the hypothesis that the two samples come from the same population, $\frac{1}{4}$ of the second sample, or m elements will be expected to fall in each interval. Let the number in the second sample which actually fall in each interval be m_1, m_2, m_3, and m_4 respectively. The test function here proposed is,

$$C = \frac{(m_1 - m)^2 + (m_2 - m)^2 + (m_3 - m)^2 + (m_4 - m)^2}{9m^2}, \tag{1}$$

where $9m^2$ is a constant, which forces C to lie on the interval 0 to 1. If the m_i, $(i = 1, 2, 3, 4)$, have values quite different from their expected value m, it is apparent that C will be large. Therefore the greater the value of C the more doubtful is the hypothesis that the two samples come from the same population. Significance values of C will be computed for several sample sizes. The question of whether C is the "best" four-interval criterion for testing the hypothesis that two samples come from the same continuous distribution is an open one

which would depend for its answer on an extensive power function analysis. We shall not go into this analysis, however, but shall use C on intuitive grounds. This case will be referred to as the case of four intervals. The extension of the method of the case of four intervals to any number of intervals presents no new difficulties in derivation, however we shall confine our attention to the cases of two and four intervals

2. The case of two intervals. Suppose $f(x)$ is a continuous distribution function with probability element $f(x)\,dx$. Let us draw a sample of size $2n + 1$ from a population having this probability element. Let the elements in the sample be $x_1, x_2, \cdots, x_{2n+1}$ ordered from least to greatest The median of this sample will be x_{n+1}. Now consider a second sample of size $2m$, and let m_1 be the number of observations, whose values are less than x_{n+1}. We call $m_2 = 2m - m_1$ the number of elements in the second sample greater than x_{n+1}

Let $p = \int_{-\infty}^{x_{n+1}} f(x)\,dx$ be the probability of an observation having a value less than x_{n+1}. Then the probability of an element having a value greater than x_{n+1} is $(1 - p)$. Thus we have the relation $f(x_{n+1})\,dx_{n+1} = dp$. The probability law of the median, x_{n+1} given by the multinomial law[1] is

$$P_r(x_{n+1}) = \frac{(2n+1)!}{n!\,1!\,n!}\,p^n(1-p)^n\,dp. \tag{2}$$

The conditional probability law of m_1, given x_{n+1}, is then

$$P_r(m_1 \mid x_{n+1}) = \frac{(2m)!}{m_1!(2m-m_1)!}\,p^{m_1}(1-p)^{2m-m_1}. \tag{3}$$

From this it follows that the joint probability law of x_{n+1} and m_1 is the product of (2) and (3) or

$$P_r(m_1, x_{n+1}) = \frac{(2n+1)!\,(2m)!}{n!\,n!\,m_1!(2m-m_1)!}\,p^{n+m_1}(1-p)^{n+2m-m_1}\,dp. \tag{4}$$

We may integrate (4) with respect to p from 0 to 1 as a Beta Function, leaving the distribution function of m_1 independent of the population probability element $f(x)\,dx$. We get for the distribution of m_1,

---

[1] The multinomial law may be stated briefly as follows.

If a trial results in one and only one of the mutually exclusive events $E_1, E_2, \cdots, E_k$, the probability P that in a total of n trials, n_1 will result in E_1, n_2 in E_2, $\cdots$, n_k in E_k, $\left(\sum_1^k n_i = n\right)$, is given by

$$P = \frac{n!}{n_1!\,n_2!\cdots n_k!}\,p_1^{n_1}p_2^{n_2}\cdots p_k^{n_k}$$

where $p_1, p_2, \cdots, p_k$, $\left(\sum_1^k p_i = 1\right)$, are the probabilities of a single trial resulting in $E_1, E_2, \cdots, E_k$ respectively.

$$(5) \qquad P_r(m_1) = \frac{(2n+1)!(2m)!(n+m_1)!(n+2m-m_1)!}{n!n!m_1!(2m-m_1)!(2n+1+2m)!}.$$

From (5) a simple recursion relation between $P_r(m_1)$ and $P_r(m_1+1)$ may be determined from which the probabilities of various values of m may be rapidly computed. For large samples it can be shown that under certain regularity conditions, the ratio, $[m_1 - E(m_1)]/\sigma_{m_1}$ may be approximated by the normal distribution[2] with zero mean and unit variance. The derivation is similar to that of the four-interval case, which is taken up in greater detail. It will be found by the use of (4) that the expected value of m_1 is m, and the variance of m_1 is $m + \frac{m(2m-1)(n+2)}{2n+3} - m^2$. Using this information, values of m_1 for various

TABLE I

The Case of Two Intervals

Lower and upper .01 and .05 percentage points for the distribution of m_1

Sample sizes		Critical values of m_1			
		Lower		Upper	
First $2n+1$	Second $2m$	$m_{1(.01)}$	$m_{1(.05)}$	$m_{1(.05)}$	$m_{1(.01)}$
11	10		1	9	
41	40	10	12	28	30
101	100	34	38	62	66
101	200	72	80	120	128
201	200	77	84	116	123
201	400	160	181	219	240
401	400	167	177	223	233
401	800	353	367	433	447
1001	1000	448	463	537	552

significance levels may be computed. The .01 and .05 percentage points of m_1 for several sample sizes are given in Table I. The values for sample sizes of 10 and 40 are computed directly from the probability law, while the larger samples have limits computed by the normal approximation. Thus for two samples of size 101 and 100, respectively, a value of m_1 less than 38 would be significant at the .05 level. Similarly, at the upper .05 level, the hypothesis would be rejected if a value of m_1 were obtained which was greater than 62. The necessity for the upper limits could easily be eliminated by testing with respect to the smaller of m_1 and m_2. However, for completeness, the upper percentage points

[2] This statement may be proved by showing that as $m, n \to \infty$ such that $m/n = constant$, the limit of the moment generating function for the ratio is identical with the moment generating function of the normal distribution with zero mean and unit variance.

are included to show the range of values of m_1 in which the hypothesis that the two samples come from the same population may be accepted.

3. The case of four intervals. If we let the first sample of size $4n + 3$ be designated by $(x_1, x_2, \cdots x_{4n+3})$, assumed drawn from a population with probability element $f(x)\,dx$ and ordered from least to greatest, then the range of x may be divided into four intervals by x_{n+1}, x_{2n+2}, and x_{3n+3}. The probability element of x_{n+1}, x_{2n+2}, x_{3n+3} is

$$\frac{(4n+3)!}{n!n!n!n!}\left(\int_{-\infty}^{x_{n+1}} f(x)\,dx\right)^n\left(\int_{x_{n+1}}^{x_{2n+2}} f(x)\,dx\right)^n\left(\int_{x_{2n+2}}^{x_{3n+3}} f(x)\,dx\right)^n\left(\int_{x_{3n+3}}^{\infty} f(x)\,dx\right)^n$$
$$\cdot f(x_{n+1})\,dx_{n+1} f(x_{2n+2})\,dx_{2n+2} f(x_{3n+3})\,dx_{3n+3}.$$

TABLE II

The Case of Four Intervals

.95 and .99 percentage points for the distribution of C

Sample sizes				$C_{.95}$	$C_{.99}$
First	Second				
$4n+3$	$4m$	n	m		
15	12	3	3	.446	.582
63	60	15	15	.113	.161
103	100	25	25	072	.102

Let

$$\int_{-\infty}^{x_{n+1}} f(x)\,dx = p_1,\quad \int_{x_{n+1}}^{x_{2n+2}} f(x)\,dx = p_2,\quad \int_{x_{2n+2}}^{x_{3n+3}} f(x)\,dx = p_3,\quad \int_{x_{3n+3}}^{\infty} f(x)\,dx = p_4.$$

The probability element of p_1, p_2, p_3, and p_4 is

$$p_r(x_{i(n+1)}) = \frac{(4n+3)!}{n!1!n!1!n!1!n!}\,p_1^n p_2^n p_3^n p_4^n\,dp_1\,dp_2\,dp_3. \tag{6}$$

Now let us consider the second sample, $(x_1', x_2', \cdots x_{4m}')$, of size $4m$. Let the number of observations falling in each of the preassigned intervals be m_i, $(i = 1, 2, 3, 4)$, where $m_4 = 4m - m_1 - m_2 - m_3$. The conditional probability of the m_i, given the values of $x_{i(n+1)}$ is also determined by the multinomial law.

$$P_r(m_i \mid x_{i(n+1)}) = \frac{(4m)!}{m_1!m_2!m_3!m_4!}\,p_1^{m_1}p_2^{m_2}p_3^{m_3}p_4^{m_4}. \tag{7}$$

The joint distribution of the p_i and the m_i is then

$$P_r(x_{i(n+1)}, m_i) = \frac{(4n+3)!\,(4m)!}{(n!)^4 m_1!m_2!m_3!m_4!}\,p_1^{n+m_1}p_2^{n+m_2}p_3^{n+m_3}p_4^{n+m_4}\,dp_1\,dp_2\,dp_3. \tag{8}$$

To obtain the distribution of the m_i alone, the p_i will be integrated out by the Dirichlet Integral[3] formula, giving a distribution which is clearly independent of the population distribution function $f(x)$.

$$(9) \quad P_r(m_i) = \frac{(4n+3)!\,(4m)!\,(n+m_1)!\,(n+m_2)!\,(n+m_3)!\,(n+m_4)!}{(n!)^4 m_1!\,m_2!\,m_3!\,m_4!\,(4m+4n+3)!}.$$

To find the expected value of the m_i, the probability law of m_1 will first be derived. The probability function for the value of x_{n+1} is

$$(10) \quad P_r(x_{n+1}) = \frac{(4n+3)!}{1!\,n!\,(3n+2)!}\, p_1^n (1-p_1)^{3n+2}\, dp_1 .$$

Then we have the conditional probability

$$(11) \quad P_r(m_1 \mid x_{n+1}) = \frac{(4m)!}{m_1!\,(4m-m_1)!}\, p_1^{m_1} (1-p_1)^{4m-m_1},$$

and

$$(12) \quad P_r(x_{n+1}, m_1) = \frac{(4n+3)!\,(4m)!}{n!\,(3n+2)!\,m_1!\,(4m-m_1)!}\, p_1^{n+m_1} (1-p_1)^{3n+2+4m-m_1}\, dp_1 .$$

To obtain the expected value of m_1, the joint distribution of m_1 and p_1 is multiplied by m_1, summed on m_1 from 0 to $4m$, and integrated on p_1 from 0 to 1.

$$(13) \quad E(m_1) = \frac{(4n+3)!}{n!\,(3n+2)!} \int_0^1 p_1^n (1-p_1)^{3n+2} \left[\sum_0^{4m} m_1 \frac{(4m)!}{m_1!\,(4m-m_1)!}\, p_1^{m_1} (1-p_1)^{4m-m_1} \right] dp_1 .$$

This interchange of the order of integration and summation is clearly valid. The quantity in brackets will be recognized as the first moment of the binomial distribution, $(p_1+q)^{4m}$ where $q = 1 - p_1$. Therefore we have

$$(14) \quad E(m_1) = \int_0^1 4m p_1 f(p_1)\, dp_1 = 4m E(p_1).$$

$E(p_1)$ and the higher moments of p_1 are found in the usual way by integrating the distributions as Beta Functions. From this we see that the expected value of m_1 is m. By repeating these operations on m_2, m_3, and m_4, it can be seen that $E(m_i) = m$, which also validates the statement made in the introduction.

---

[3] A discussion of the Dirichlet Integral may be found in Woods—*Advanced Calculus*, p 167 It may be stated as follows for the problem in which we are interested

$$\iiint x^{l-1} y^{m-1} z^{n-1} (1-x-y-z)^{r-1}\, dx\, dy\, dz = \frac{\Gamma(l)\Gamma(m)\Gamma(n)\Gamma(r)}{\Gamma(l+m+n+r)},$$

where we integrate over the region bounded by $x + y + z = 1$, and the three coordinate planes.

We have previously presented the criterion (1).

The next problem is to find a distribution function to which the distribution of C may be fitted. A reasonable choice appears to be the Pearson Type I curve.

$$f(x) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} x^{r-1}(1-x)^{s-1}. \tag{15}$$

The distribution of C is fitted by equating the first two moments of the two distributions and solving for the constants r and s of the Type I distribution. Using the theorem that the mean value of the sum of variates is equal to the sum of their mean values, we have

$$E(C) = \frac{1}{9m^2}[E(m_1^2) + E(m_2^2) + E(m_3^2) + E(m_4^2) - 4m^2]. \tag{16}$$

Also the second moment may be written as

$$\begin{aligned} E(C^2) = \frac{1}{81m^4}[&E(m_1^4) + E(m_2^4) + E(m_3^4) + E(m_4^4) + 16m^4 + 2E(m_1^2 m_2^2) \\ &+ 2E(m_1^2 m_3^2) + 2E(m_1^2 m_4^2) + 2E(m_2^2 m_3^2) + 2E(m_2^2 m_4^2) \\ &+ 2E(m_3^2 m_4^2) - 8m^2\{E(m_1^2) + E(m_2^2) + E(m_3^2) + E(m_4^2)\}]. \end{aligned} \tag{17}$$

The expected value of m_1^2 is found in the same manner as $E(m_1)$ and here also it can be shown that the $E(m_i^2)$ are all equal. The same procedure holds for $E(m_i^4)$.

$$\begin{aligned} E(m_i^2) &= m + \frac{m(4m-1)(n+2)}{4n+5}, \\ E(m_i^4) &= m + \frac{7m(4m-1)(n+2)}{4n+5} + \frac{6m(4m-1)(4m-2)(n+3)(n+2)}{(4n+6)(4n+5)} \\ &\quad + \frac{m(4m-1)(4m-2)(4m-3)(n+4)(n+3)(n+2)}{(4n+7)(4n+6)(4n+5)}. \end{aligned} \tag{18}$$

By using the moment generating function of the trinomial distribution, the $E(m_i^2 m_j^2)$ may also be found in a similar manner

$$\begin{aligned} E(m_i^2 m_j^2) = &\frac{m(4m-1)(n+1)}{4n+5} + \frac{2m(4m-1)(4m-2)(n+1)(n+2)}{(4n+6)(4n+5)} \\ &+ \frac{m(4m-1)(4m-2)(4m-3)(n+2)(n+1)(n+2)}{(4n+7)(4n+6)(4n+5)}. \end{aligned} \tag{19}$$

As a result we have

$$E(C) = \frac{4}{9m} + \frac{4(4m-1)(n+2)}{9m(4n+5)}. \tag{20}$$

Let $E(C) = A$ to simplify later relations to be computed. Finally

$$E(C^2) = \frac{4}{81m^3}\Bigg[1 + \frac{7(4m-1)(n+2)}{4n+5} + \frac{6(4m-1)(4m-2)(n+3)(n+2)}{(4n+6)(4n+5)} + \frac{(4m-1)(4m-2)(4m-3)(n+4)(n+3)(n+2)}{(4n+7)(4n+6)(4n+5)} + 4m^3 + \frac{3(4m-1)(n+1)}{4n+5} + \frac{6(4m-1)(4m-2)(n+1)(n+2)}{(4n+6)(4n+5)} + \frac{3(4m-1)(4m-2)(4m-3)(n+2)^2(n+1)}{(4n+7)(4n+6)(4n+5)} - 8m^2 - \frac{8m^2(4m-1)(n+2)}{4n+5}\Bigg]. \tag{21}$$

To simplify later relations we let $E(C^2) = B$.

The first two moments of the Type I distribution are easily found to be

$$\mu_1 = \frac{r}{r+s} = A \qquad \mu_2 = \frac{\mu_1(r+1)}{(r+s+1)} = B. \tag{22}$$

Solving these two simultaneous equations for r and s,

$$r = \frac{B-A}{A-\frac{B}{A}} \qquad s = \frac{r}{A} - r. \tag{23}$$

A number of percentage points for the Type I distribution have been computed by Miss Catherine Thompson, [3]. Using these limits, the hypothesis may be accepted or rejected as to whether or not the two samples come from the same population.

Table II shows the .95 and .99 percentage points of C for three sample sizes.

4. Summary. The problem considered here is that of devising a simple method of testing the hypothesis that two samples are from identical populations having continuous distribution functions It may be summarized briefly as follows. The first sample is used to establish any desired number of intervals into which the observations of the second sample may fall A test criterion is proposed which is based on the deviations of the numbers of elements of the second sample which fall in the intervals from the expected values of the respective numbers. Two cases are discussed, that of two intervals and that of four intervals, making use of the median and quartiles in the first sample to determine the intervals. Tables of 1% and 5% points for several sample sizes of both cases are given.

REFERENCES

[1] A. Wald and J. Wolfowitz, *Annals of Math. Stat*, Vol. 11 (1940), p. 147.
[2] W. J. Dixon, *Annals of Math. Stat*, Vol 11 (1940), p. 199.
[3] Catherine Thompson, *Biometrika*, Vol. 32 (1941), p. 151.

NOTES

This section is devoted to brief research and expository articles, notes on methodology and other short items.

NOTE ON THE INDEPENDENCE OF CERTAIN QUADRATIC FORMS

By Allen T. Craig

University of Iowa

Various approaches to the problem of the independence of quadratic forms in normally and independently distributed variables have been made by R. A. Fisher, Cochran, Madow and others. It is the purpose of this note to point out a few simple propositions which, in so far as the writer is aware, have not had specific mention in the literature.

1. Independence of certain quadratic forms. Theorem 1: *A necessary and sufficient condition that two real symmetric quadratic forms, in n normally and independently distributed variables, be independent in the probability sense is that the product of the matrices of the forms be zero.*

Let the chance variable x be normally distributed with mean zero and unit variance. Let $x_1, x_2, \cdots, x_n$ be n independent values of x and let A and B be two real symmetric matrices, each of order n. Write $Q_1 = \Sigma\Sigma a_{ij}x_ix_j$ and $Q_2 = \Sigma\Sigma b_{ij}x_ix_j$, where $\| a_{ij} \| = A$ and $\| b_{ij} \| = B$. It is well known that the generating function of the moments of the joint distribution of Q_1 and Q_2 can be written

$$G(\lambda, \lambda') = | I - \lambda A - \lambda' B |^{-\frac{1}{2}},$$

so that

$$| I - \lambda A - \lambda' B | = | I - \lambda A | \, | I - \lambda' B |, \tag{1}$$

for all real values of λ and λ', is necessary and sufficient for the independence of Q_1 and Q_2.

If Q_1 and Q_2 are independent, then (1), being true for all real values of λ and λ', is in particular true for $\lambda = \lambda'$. Thus

$$| I - \lambda(A + B) | \equiv | I - \lambda A | \, | I - \lambda B |. \tag{2}$$

Denote by r_1, r_2 and $r \leq r_1 + r_2$ respectively the ranks of A, B and $A + B$. Then $r = r_1 + r_2$ since (2) expresses the identity of two polynomials in λ of degrees r and $r_1 + r_2$.

Further, if we write

$$| I - \lambda A | = (1 - \lambda p_1) \cdots (1 - \lambda p_{r_1}),$$

$$| I - \lambda B | = (1 - \lambda q_1) \cdots (1 - \lambda q_{r_2}),$$

and $|I - \lambda(A + B)| = (1 - \lambda s_1) \cdot \cdot (1 - \lambda s_{r_1+r_2})$, then, because the factorization of polynomials is unique, each s_j can be paired with one of the numbers $p_1, \cdots, p_{r_1}, q_1, \cdots, q_{r_2}$. Thus, if Q_1 and Q_2 are independent, the rank of $A + B$ is the sum of the ranks of A and B, and the non-zero roots of the characteristic equation of $A + B$ are those of the characteristic equation of A together with those of the characteristic equation of B. There exists an appropriately chosen orthogonal matrix L of order n such that $L'(A + B)L$, L' being the conjugate of L, is a matrix with the reciprocals of the numbers $p_1, \cdots, p_{r_1}$, $q_1, \cdots, q_{r_2}$ on the principal diagonal and zeros elsewhere. Then $L'AL$ and $L'BL$ have no overlapping non-zero elements and $L'ALL'BL = 0$. But $L' = L^{-1}$, the inverse of L. Hence, upon multiplying both members of the preceding equation on the right by L' and on the left by L, we have $AB = 0$. Since $A = A'$ and $B = B'$, likewise $BA = 0$.

Conversely, suppose $AB = 0$. Then the matrix $(I - \lambda A)(I - \lambda' B) = I - \lambda A - \lambda' B$. These matrices being equal, their determinants are equal and the condition (1) for the independence of Q_1 and Q_2 is satisfied

The theorem is readily extended to the case of the mutual independence of any finite number of such quadratic forms.

The product of a non-singular matrix and a matrix of rank R is a matrix of rank R. Hence, every non-singular quadratic form of the kind here discussed is correlated with every non-identically vanishing quadratic form in the same variables.

2. Conditions for independent Chi-Square distributions. The preceding theorem enables one to determine, by multiplication of matrices, whether real symmetric quadratic forms in normally and independently distributed variables are themselves independent in the probability sense. The following theorem affords a simple test as to whether the distributions are of the Chi-Square type.

THEOREM 2: *Necessary and sufficient conditions that each of two real symmetric quadratic forms, in n normally and independently distributed variables with mean zero and unit variance, be independently distributed as is Chi-Square, are that the product of the matrices of the forms be zero and that each matrix equal its own square.*

If Q_1 and Q_2 are independently distributed as is Chi-Square, then $AB = 0$ and each of the non-zero roots of the characteristic equations of A and B is $+1$. For an appropriately chosen orthogonal matrix L, of order n, $L'AL$ is a matrix with r_1 elements on the principal diagonal $+1$, all other elements being zero. For such a matrix it is seen that $(L'AL)(L'AL) = L'A^2L = L'AL$ and $A^2 = A$. A similar argument shows that $B^2 = B$.

Conversely, if $AB = 0$, then Q_1 and Q_2 are independent Further, if $A^2 = A$ and $B^2 = B$, each of the non-zero roots of the characteristic equations of A and B is $+1$ This follows from the fact that the roots of the characteristic equation of the square of any matrix are themselves the squares of the roots of the

characteristic equation of that matrix. Since A and B are real and symmetric, the roots under consideration are real. Thus Q_1 and Q_2 have independent Chi-Square distributions with r_1 and r_2 degrees of freedom respectively.

This theorem can likewise be extended to any finite number of these quadratic forms

Of special interest is the case of, say k, quadratic forms for which the sum of the k matrices is the identity matrix Thus $A_1 + A_2 + \cdots + A_k = I$. By Theorem 1, it is both necessary and sufficient for the mutual independence of the k forms that $A_u A_v = 0$, $u \neq v$.

Now

$$A_i = I - A_1 - \cdots - A_{i-1} - A_{i+1} - \cdots - A_j - \cdots - A_k$$

and

$$A_i A_j = A_j - A_1 A_j - \cdots - A_{i-1} A_j - A_{i+1} A_j - \cdots - A_j^2 - \cdots - A_k A_j ,$$

so that $A_j = A_j^2$. In this particular case it is to be seen that the mutual independence of the forms implies that their several distributions are of the Chi-Square type.

A CHARACTERIZATION OF THE NORMAL DISTRIBUTION

By Irving Kaplansky

Harvard University

In 1925 R. A. Fisher gave a geometric derivation of the joint distribution of mean and variance in samples from a normal population (*Metron*, Vol. 5, pp. 90–104). On examining the argument however, we find that an (apparently) more general result is actually established: if $f(x_1) \cdots f(x_n)$ is a function $g(m, s)$ of the sample mean m and standard deviation s, then the probability density of m and s in samples of n from the population $f(x)$ is $g(m, s)s^{n-2}$. This condition on $f(x)$ is of course satisfied if $f(x)$ is normal; in this note we shall conversely show that for $n \geqq 3$ it characterizes the normal distribution. In the proof it will be assumed that $g(m, s)$ possesses partial derivatives of the first order, although a weaker assumption would probably suffice.

Let us for the moment restrict the variables x_i to values such that $f(x_i) > 0$. After a change of notation we have

$$\phi(x_1) + \cdots + \phi(x_n) = h(u, v),$$

where $\phi = \log f$, $u = x_1 + \cdots + x_n$, $v = \frac{1}{2}(x_1^2 + \cdots + x_n^2)$. A differentiation yields

$$\phi'(x_i) = h_u + h_v x_i .$$

Solving two of these equations for h_v, we find

$$h_v = \frac{\phi'(x_i) - \phi'(x_j)}{x_i - x_j}, \qquad (i \neq j), \tag{1}$$

and, for $n \geqq 3$, it follows that the right member of (1) is a constant, say 2A. Then

$$\phi'(x_i) - 2Ax_i = \phi'(x_j) - 2Ax_j = \text{a constant } B.$$

$$\phi(x) = Ax^2 + Bx + C.$$

We now have $f(x) = e^{\phi(x)}$ whenever $f(x) > 0$; but since $f(x)$ is continuous, this implies $f(x) = e^{\phi(x)}$ everywhere.

NEWS AND NOTICES

Readers are invited to submit to the Secretary of the Institute news items of general interest

Personal Items

Dr. Holbrook Working has been appointed Chief Statistical Consultant on Industrial Processes and Products in the Office of Production Research and Development of the War Production Board.

Professor Harold Hotelling of Columbia University was the official representative of the Institute of Mathematical Statistics at the Copernican Quadricentennial Celebration which was held in New York City on May 24.

Dr Edward B. Olds has taken a position with the Curtiss-Wright Corporation.

Dr. Nilan Norris is a Sergeant with the Fourth Statistical Control Unit of the Fourth Air Force with headquarters at San Francisco, California.

Dr Edward Helly is with the Signal Corps Training Program at Illinois Institute of Technology.

Dr. C. W. Cotterman is in the United States Army at Camp Grant, Illinois.

Mr. M D. Bingham has been commissioned an Ensign in the United States Naval Reserve and is stationed at Fort Schuyler, New York.

Lt. George W. Petrie, USNR, is teaching in the Midshipmen's School at Notre Dame, Ind.

New Members

The following persons have been elected to membership in the Institute:

Arias B., Jorge Civ. Eng (Guatemala) Eng., Rural Electrification Administration, 420 Locust St., St. Louis, Mo

Bailey, A. L. B.S. (Michigan) Stat., American Mutual Alliance, 60 East 42 St., New York, N. Y.

Becker, Harold W. Instr , Mare Island Trainee School. *126 Benson Ave., Vallejo, Calif.*

Bernstein, Shirley R. B.S. (Carnegie Inst. Tech) Res Asst., United Steelworkers of America, Pittsburgh, Pa. *5501 Beverly Pl.*

Bickerstaff, Asst. Prof. Thomas A. M.A (Mississippi) Univ. of Miss., University, Miss.

Birnbaum, Asst. Prof. Z. William. Ph D. (Lwow) Univ. of Wash., Seattle, Wash.

Brumbaugh, Prof. Martin A. Ph.D (Pennsylvania) Univ. of Buffalo, Buffalo, N. Y.

Burrows, Glenn L. B.A. (Michigan State Coll.) Instr., Michigan State Coll., East Lansing, Mich

Cohen, Jozef B. B S. (Chicago) Sage Fellow in Psychology, Cornell Univ., Ithaca, N. Y.

Cope, Asso. Prof. T. Freeman. Ph.D. (Chicago) Queens College, Flushing, N. Y.

Cudmore, Sedley A. M.A (Oxford) Stat., Dominion Bur. of Stat., Ottawa, Canada.

Cureton, Edward E. Ph D. (Columbia) Sr. Personnel Technician, War Dept., *RFD 1, Tauxemont, Alexandria, Va.*

De Castro, Prof. Lauro S. V. Civ. Eng. (Escola Nacional de Enginharia) Catholic Univ., Rio de Janeiro, Brazil *62 rua David Campista.*

Edwards, G. D. A B (Harvard) Dir. of Quality Assurance, Bell Telephone Laboratories, 463 West St , New York, N Y.

Gifford, Kenneth R. Student, Mass. Inst. Tech., Cambridge, Mass. *97 Bay State Rd., Boston, Mass*

Gottfried, Bert A. A.M. (Columbia) Stat. Clerk, 4300 Kaywood Dr., Mt. Ranier, Md.
Hamilton, Prof. Thomas R. Ph D. (Columbia) Texas A & M Coll., College Station, Tex.
Heide, J. D. M.S (Iowa) Stat., U S Rubber Co., *1324 Altoona Ave , Eau Claire, Wisc.*
Hilfer, Irma. M A. (Columbia) Actuary, N. Y. C. Board of Transportation, *165 W. 97 St., New York, N. Y.*
Howell, John M. B.A. (UCLA) Stat., Northrop Aircraft Inc., Hawthorne, Calif. *4140 W. 63 St., Los Angeles, Calif*
Hurwicz, Leonid. L.L M. (Warsaw) Res. Asso , Cowles Comm., Univ of Chicago, Chicago, Ill
Kendall, Maurice G. M A (Cambridge) Stat., Chamber of Shipping of the United Kingdom, *Richmond House, Aldenham Rd , Bushey, Eng*
Klein, Lawrence R. B.A (California) Teaching Fellow, Mass Inst Tech., Cambridge, Mass.
Kuznets, George M. Ph.D (California) Instr., Giannini Foundation, Univ of Calif., Berkeley, Calif.
Landau, H. G. M.S. (Carnegie Inst Tech) Stat. Analyst, War Dept., Washington, D. C *2408 20 St., N.E*
Langmuir, Charles R. Ed M. (Harvard) Carnegie Foundation. *437 West 59 St., New York, N. Y.*
Levy, Henry C. L L.B. (Fordham) Instr., N Y C C., New York, N. Y *600 West 116 St.*
Li, Jerome C. R. B S. (Nanking) Student, Iowa State Coll , Ames, Iowa. *2184 Lincoln Way.*
Lieberman, Jacob E. B S. (Brooklyn Coll.) Jr Stat , Census Bureau, Washington, D. C. *2422 14 St., N E.*
Martin, Margaret P. M A (Minnesota) Instr., Columbia Univ , New York, N Y. *1230 Amsterdam Ave.*
Nash, Stanley W. B A. (Coll of Puget Sound) San Joaquin Experimental Range, O'Neals, Calif.
Norton, Horace W. Ph.D. (London) Sr Meteorologist, U S. Weather Bur., Washington, D C *3118 North First Rd , Arlington, Va.*
Olds, Edward B. Ph.D (Pittsburgh) Stat., Curtiss-Wright Corp. *298 Niagra Falls Blvd., Buffalo, N. Y.*
Preston, Bernard. C P A , *103 Park Ave., New York, N Y.*
Rosenblatt, David. B.S (Coll. City of N Y) Asst Stat., *1422 Whittier St , N W , Washington, D C.*
Sard, Asst. Prof. Arthur. Ph.D (Harvard) Queens College, Flushing, N. Y. *146–19 Beech Ave.*
Schapiro, Anne. B.A. (Bryn Mawr) Jr. Analyst, Institute of Applied Econometrics, *350 W. 57 St., New York, N. Y.*
Simpson, William B. Grad. Student, Columbia Univ , New York, N. Y
Springer, Melvin D. M.S (Illinois) Asst Instr , Univ of Illinois, Urbana, Ill.
Stein, Irving. B.S (Mass Inst Tech) Asso. Stat , War Dept., Washington, D C. *611 Oglethorpe St.*
Stergion, Andrew P. M.S (Mass Inst Tech.) 1st Lt , USA, The Proving Center, Aberdeen Proving Gd., Md.
Sternhell, Arthur I. B.A. (New York) Staff Asst , Metropolitan Life Ins. Co., *1938 E. Tremont Ave., Parkchester, N. Y.*
Thompson, Louis T. E. Ph.D. (Clark) Dir Res. and Dev., Lukas-Harold Corp., Indianapolis, Ind. *340 East Maple Rd.*
Tyler, Asst. Prof. George W. M.A (Duke) Virginia Polytechnic Inst , Blacksburg, Va.
Working, Holbrook S. Ph D. (Wisconsin) Chief Stat. Consultant, War Production Board, Washington, D C. *Food Res. Inst., Stanford Univ , Calif*

The following persons have been elected to Junior membership in the Institute.

Blumenthal, Lydia. Hunter College, New York, N. Y. *1001 Lincoln Pl., Brooklyn, N. Y.*

Gunlogson, Lee. Univ of Minnesota, Minneapolis, Minn. *1906 Third Ave.*

Heacock, Richard R. Oregon State Coll , Corvallis, Ore. *P O. Box 207, Seaside, Ore.*

Locatelli, Humbert J. Columbia Univ., New York, N. Y. *44 Seaman Ave.*

Mathisen, Harold C., Jr. Princeton Univ., Princeton, N J *4 Middle Dod Hall.*

Murphy, Ray Bradford. Princeton Univ , Princeton, N J. *28 Godfrey Rd., Upper Montclair, N. J.*

Peters, Edward J., Jr. Georgetown Univ., Washington, D. C. *126 St. James Pl., Atlantic City, N. J*

Smith, Joan T. Univ of Minnesota, St. Paul, Minn. *673 East Nebraska Ave.*

SPECIAL COURSES IN STATISTICAL QUALITY CONTROL

The application of statistics to quality control is now being furthered in a program in which the War Production Board and the U. S. Office of Education are cooperating to assist statisticians in various industrial areas to provide suitable courses of instruction sponsored by their own institutions.

The general plan of the program has been influenced by two conclusions drawn from the experience gained in ESMWT courses carried on by Stanford University during 1942–43.[1] These conclusions were: (1) that a short full-time course in statistical quality control tends to be peculiarly effective; and (2) that it is vital to have the initial courses followed by meetings in which the course members gather to report on applications they have made and to receive encouragement and any needed assistance.

The giving of short full-time courses presents a problem of assembling a suitable staff, since four instructors will ordinarily be needed. If this problem were solved by arranging for a single staff to tour all the principal industrial regions giving courses in quality control, the local leadership necessary for establishing widespread use of statistical methods of quality control in industry would not be developed. The program adopted seems to offer an effective solution of these problems.

Under the program now in effect, the War Production Board, through its Office of Production Research and Development supplies an experienced person to assist with the arrangement of courses and to participate in the instruction.[2] Two of the instructors in each course will ordinarily be provided by a local educational institution, which will also promote the course and make necessary local arrangements through its institutional representative of the Engineering Science and Management War Training program. It is not considered necessary that the instructors provided by the institution have previous experience with statistical quality control provided they are sufficiently competent in the theory of sampling, but it is desirable that at least one of them have practical experience with quality control. It may often happen that one of the instructors can be a quality control man from a local industrial establishment. The representative of the WPB will assist with arrangements for bringing in one (or, where needed, two) additional outside instructors.

The sponsoring institution costs for the courses, which do not include the salary and expenses of the representative of the WPB, may be provided through the ESMWT program. The follow-up work with men who have taken the initial courses may be arranged also as part of the ESMWT program of the

---

[1] A description of these courses offered by Stanford University appeared in the *Annals of Mathematical Statistics*, March 1943, p. 96.

[2] At present Professor Holbrook Working is serving in this capacity

educational institution sponsoring the original course. The follow-up work should be handled by a local instructor who participated in the original course.

The two basic courses and the one follow-up course that have already been given by Stanford University were conducted under essentially the plan outlined above, except that they did not have the benefit of assistance from the WPB. Three courses have thus far (May 25) been arranged under the new plan: one sponsored by Rhode Island State College, to be held during May 27 to June 2 at Newport, and two sponsored by Stanford University, to be held respectively in Los Angeles, June 13 to 20, and in San Francisco, June 22 to 29. Preliminary steps have been taken toward the arrangement of several additional courses.

REPORT OF THE NEW YORK MEETING OF THE INSTITUTE

A joint meeting between the Institute and the American Society of Mechanical Engineers was held on Saturday, May 29, 1943 at the Engineering Societies Building, 29 West 39th Street, New York City. Of the ninety-five individuals attending the meeting, the following fifty-seven members of the Institute were present:

Theodore W. Anderson, K. J. Arnold, Robert E. Bechhofer, B. M. Bennett, C. I. Bliss, Mary E Boozer, P Boschan, A H. Bowker, Burton H. Camp, A. C. Cohen Jr., H. F. Dodge, C. Eisenhart, Mary L Elveback, W. C. Flaherty, H. Goode, John I. Griffin, Charles C. Grove, Frank E. Grubbs, E. J. Gumbel, Harold Hotelling, J. M Juran, B. F. Kimball, Lila Knudsen, Howard Levene, E. Vernon Lewis, Simon Lopata, Frank W. Lynch, Henry Mann, E. C Molina, N Morrison, Philip J. McCarthy, Luis F. Nanni, Franklin S. Nelson, M. L. Norden, P. S. Olmstead, R F. Passano, Edward Paulson, G. A D. Preinreich, A C. Rosander, Arthur Sard, Henry Scheffé, Bernice Scherl, Edward M. Schrock, L W. Shaw, William B. Simpson, S G. Small, Arthur Stein, Andrew P. Stergion, M. Stevens, David F Votaw Jr., A Wald, Helen M Walker, W. A. Wallis, S S. Wilks, J. Wolfowitz, L C. Young

The general topic of the meeting was *Industrial Applications of Statistics.* At the morning session the following papers were presented, with Professor Harold Hotelling presiding:

1 *On the Theory of Runs with some Application to Quality Control*
J. Wolfowitz.
2. *On the Presentation of Data as Evidence.*
Churchill Eisenhart.

At the afternoon session, the following papers were presented with Mr. E. C. Molina as Chairman:

1. *A Sampling Inspection Plan for Continuous Production.*
H F. Dodge
2 *Tolerances and Product Acceptability*
L. C Young.

A meeting of the Board of Directors was held after the afternoon session.

EDWIN G. OLDS
Secretary

Photo Lithoprint Reproduction
EDWARDS BROTHERS, INC
Lithoprinters
ANN ARBOR MICHIGAN
1947

THE ANNALS *of* MATHEMATICAL STATISTICS

(FOUNDED BY H. C. CARVER)

THE OFFICIAL JOURNAL OF THE INSTITUTE OF MATHEMATICAL STATISTICS

Contents

	PAGE
The Comparison of Different Scales of Measurement for Experimental Results. W. G. COCHRAN	205
On Stochastic Limit and Order Relationships. H. B. MANN AND A. WALD	217
On a Measure Problem Arising in the Theory of Non-Parametric Tests. HENRY SCHEFFÉ	227
Further Results on Probabilities of a Finite Number of Events. KAI LAI CHUNG	234
On the Problem of Testing Hypotheses. R. v. MISES	238
On the Reliability of the Classical Chi-Square Test. E. J. GUMBEL	253
A Sampling Inspection Plan for Continuous Production. H. F. DODGE	264
On the Theory of Runs with Some Applications to Quality Control. J. WOLFOWITZ	280
The Accuracy of Sampling Methods in Ecology. PAUL G. HOEL	289
News and Notices	301
Report on the Washington Meeting of the Institute	302
Report on the First Meeting of the Pittsburgh Chapter of the Institute	303

Vol. XIV, No. 3 — September, 1943

THE ANNALS OF MATHEMATICAL STATISTICS

EDITED BY

S. S. WILKS, *Editor*

A. T. CRAIG
W. E. DEMING
T. C. FRY
H. HOTELLING
J. NEYMAN
W. A. SHEWHART

WITH THE COÖPERATION OF

W. G. COCHRAN
J. H. CURTISS
H. F. DODGE
P. S. DWYER
C. EISENHART
W. K. FELLER
P. G. HOEL
W. G. MADOW
A. WALD

The ANNALS OF MATHEMATICAL STATISTICS is published quarterly by the Institute of Mathematical Statistics, Mt. Royal & Guilford Aves., Baltimore, Md. Subscriptions, renewals, orders for back numbers and other business communications should be sent to the ANNALS OF MATHEMATICAL STATISTICS, Mt. Royal & Guilford Aves., Baltimore, Md., or to the Secretary of the Institute of Mathematical Statistics, E. G. Olds, Carnegie Institute of Technology, Pittsburgh, Pa. Changes in mailing address which are to become effective for a given issue should be reported to the Secretary on or before the 15th of the month preceding the month of that issue. The months of issue are March, June, September and December. Because of war-time difficulties of publication, issues may often be from two to four weeks late in appearing. *Subscribers are therefore requested to wait at least 30 days after month of issue before making inquiries concerning non-delivery.*

Manuscripts for publication in the ANNALS OF MATHEMATICAL STATISTICS should be sent to S. S. Wilks, Fine Hall, Princeton, New Jersey. Manuscripts should be typewritten double-spaced with wide margins, and the original copy should be submitted. Footnotes should be reduced to a minimum and whenever possible replaced by a bibliography at the end of the paper; formulae in footnotes should be avoided. Figures, charts, and diagrams should be drawn on plain white paper or tracing cloth in black India ink twice the size they are to be printed. Authors are requested to keep in mind typographical difficulties of complicated mathematical formulae.

Authors will ordinarily receive only galley proofs. Fifty reprints without covers will be furnished free. Additional reprints and covers furnished at cost.

The subscription price for the ANNALS is $5.00 per year. Single copies $1.50. Back numbers are available at $5.00 per volume, or $1.50 per single issue.

COMPOSED AND PRINTED AT THE
WAVERLY PRESS, INC.
BALTIMORE, MD., U. S. A.

Entered as second-class matter at the Post Office at Baltimore, Maryland, under the Act of March 3, 1879

THE COMPARISON OF DIFFERENT SCALES OF MEASUREMENT FOR EXPERIMENTAL RESULTS[1,2]

By W. G. Cochran

Iowa State College

1. Introduction. In some fields of research, the development of a satisfactory method for measuring the effects of experimental treatments constitutes a difficult problem. The estimation of the vitamin content of preparations of foods furnishes a good example, for most of the vitamins several years of work were required to construct a reliable method of assay. In other cases, where the ideal method for measuring treatment responses is costly or troublesome, a search may be made for a more convenient substitute Thus in pasture or forage-crop experiments the species composition of a plot may be estimated by eye inspection as a substitute for a complete botanical separation. As a third example we may quote experiments in cookery, where the flavor and quality of the dishes are subject to the whims of human taste. Frequently a panel of judges is employed, each of whom scores the dishes independently It is not easy to determine how the panel should be chosen, nor how representative its verdicts are of consumer preferences in general

When such problems are investigated, experiments may be carried out specifically for the purpose of comparing two or more methods or *scales* of measurement. Where the process of measurement affects only the final stages of the experiment, as in the last two examples quoted above, all that is necessary is to score the *same* experiment by the various scales under consideration. In comparing two different methods of assaying vitamins, on the other hand, independent experiments are frequently required, the only common feature being that the same set of treatments is tested in both experiments.

In the interpretation of the results of such experiments, two types of comparison are of general interest One concerns the *relations* between the scales. It may be summed up rather loosely in the question· Are the effects of the treatments the same in all scales? For a more exact formulation, consider the case of two scales, which is probably the most frequent in practice. Let ξ_{1t}, ξ_{2t} be the true means of the tth treatment as measured on the two scales. We may wish to examine the following hypotheses:

(i) *Scales equivalent.*

$$\xi_{1t} = \xi_{2t}, \qquad \text{(all } t\text{)}; \tag{1}$$

(ii) *Scales equivalent, apart from a constant difference.*

$$\xi_{1t} = \xi_{2t} + \epsilon, \qquad \text{(all } t\text{)}; \tag{2}$$

---

[1] Paper presented at a meeting of the Institute of Mathematical Statistics, Washington, D C , June 18, 1943

[2] Journal Paper No J-1136 of the Iowa Agricultural Experiment Station, Ames, Iowa. Project 514

(iii) *Scales linearly related:*

$$\alpha\xi_{1t} + \beta\xi_{2t} = \gamma, \qquad \text{(all } t\text{)}; \tag{3}$$

(iv) *Relation monotonic, but not linear:*

$$\xi_{1t} = f(\xi_{2t}, \alpha, \beta, \cdots), \qquad \text{(all } t\text{)}; \tag{4}$$

where the function is strictly monotonic.

In this case the two scales are mutually consistent in that they place any set of treatments in the same order. The ratio of a treatment difference in one scale to the corresponding difference in the other scale is, however, not constant.

(v) *Relation not monotonic.* Here the scales do not place the treatments in the same order and consequently are not satisfactory substitutes for each other.

The second question concerns the relative *accuracy* or *sensitivity* of the two scales. For practical purposes this question may be put as follows: how many replications are required with the second scale to attain the accuracy given by r replications with the first scale? It is clear that the answer depends both on the experimental errors associated with the scales and on the magnitudes of the treatment effects in the two scales. For example, Coward [1] reports that in the assay of vitamin D, male rats give a higher experimental error than females, yet provide a more accurate assay because they are more responsive The relative accuracy may be different in different parts of the two scales. This is likely to happen whenever the relation between the scales is of type (iv) above.

This paper gives a preliminary discussion of some of the simpler questions raised above, to which recent work in multivariate analysis is applicable. A complete solution for small sample work appears to demand considerable further development in the distribution theory of multivariate analysis.

The discussion is confined to the case in which all scales measure the same experiment. The case where each scale requires a separate experiment may be expected to be somewhat simpler, but cannot conveniently be treated as a special case of the procedure for a single experiment.

2. Assumptions. Let $x_1, x_2, \cdots x_p$ denote measurements on the p scales and let n_1 and n_2 be the numbers of degrees of freedom for treatments and error respectively. The experimental data furnish a joint analysis of variance and covariance of the p variates as follows:

(5)

	d.f.	*Sum of squares or products*
Mean	1	m_{ij}
Treatments	n_1	a_{ij}
Error	n_2	b_{ij}

It will be assumed that $x_1, \cdots, x_p$ follow a multivariate normal distribution, and that for any pair of variates x_i, x_j the error mean covariance σ_{ij} is constant throughout the experiment (though it may vary as i and j vary). Thus the

quantities b_{ij}, follow the standard joint distribution, Wishart [16], of sums of squares and products while the quantities m_{ij} and a_{ij} follow the corresponding non-central distributions and the three sets of distributions are independent.

3. Tests for equivalence. If there are only two scales, a test for equivalence is obtained from elementary techniques. An analysis of variance similar to (5) is computed on the *differences* between the two scales for every observation. If equations (1) hold in the population, the sums of squares for the Mean, Treatments and Error are distributed independently as $\chi^2(\sigma_{11} + \sigma_{22} - 2\sigma_{12})$. The pooled mean square for the Mean and Treatments may therefore be compared with the Error mean square in a variance-ratio test, the degrees of freedom being $(n_1 + 1)$ and n_2. If the scales are equivalent apart from a constant difference, the same result is valid for Treatments and Error, while the mean square for the Mean is proportional to a non-central χ^2. Thus separate z- or F-tests on the Mean and Treatments assist in distinguishing between hypotheses (1) and (2).

4. More than two scales. Let ξ_{it} be the true mean of the tth treatment as measured on the ith scale. The first two hypotheses may now be written respectively:

$$\xi_{it} = \xi_t \tag{1'}$$

$$\xi_{it} = \xi_t + \epsilon_i \tag{2'}$$

for $i = 1, 2, \cdots, p$. The quantities ϵ_i, whose sum may be assumed zero, measure the constant differences among the scales.

If the interactions of all components with Scales are computed, the analysis of variance extends formally, with the following separation of degrees of freedom:

(6)

	d.f.
Mean × Scales	$(p - 1)$
Treatments × Scales	$n_1(p - 1)$
Error	$n_2(p - 1)$

The three lines in the analysis play the same roles as before in relation to hypotheses (1') and (2'). When $p > 2$, however, it may be shown that the three sums of squares are not distributed as multiples of χ^2 unless (i) all scales have the same error variance and (ii) every pair of scales has the same correlation coefficient. Where these conditions are reasonably well satisfied, as happens possibly when experienced judges employ a similar scoring system, the above analysis supplies approximate tests. But with scales which differ widely in their experimental errors or in their degrees of interrcorrelation, the validity of variance-ratio tests is open to more serious question.

In order to obtain an exact test, we may note that hypothesis (1') is closely related to the Wilks-Lawley hypothesis (Wilks [15], Lawley [9], Hsu [7]) that the means of k populations are all equal. If each treatment denotes a separate population, the Wilks-Lawley hypothesis states that

$$\xi_{it} = \xi_i \qquad (t = 1, 2, \cdots, n_1 + 1). \tag{7}$$

Since this differs from (1′) only in the interchange of the letters i and t, it is clear that the two hypotheses may be subjected to the same kind of test.

For the details of the procedure we first divide the $(p-1)$ comparisons among scales into $(p-1)$ single comparisons by the introduction of a set of variates y_i, $(i = 1, 2, \cdots, p-1)$.

$$y_i = \sum_{j=1}^{p} \lambda_{ij} x_j . \tag{8}$$

Any set of y's may be chosen, provided that they are linearly independent and that

$$\sum_{j=1}^{p} \lambda_{ij} = 0, \qquad (i = 1, 2, \cdots (p-1)). \tag{9}$$

Thus with three scales we might use $y_1 = x_1 - x_2$, $y_2 = x_1 - x_3$ or $y_1 = 2x_1 - x_2 - x_3$, $y_2 = x_2 - x_3$.

The next step is to compute an analysis of variance and covariance of the y variates, as follows:

(10)

	d.f.	*Sum of squares or products*
Mean	1	m'_{ij}
Treatments	n_1	a'_{ij}
Error	n_2	b'_{ij}

If hypothesis (1′) holds, it follows from (9) that the three sets of quantities m'_{ij}, a'_{ij} and b'_{ij} all follow the standard joint distribution for sums of squares and products. Hence Wilks' test (Wilks [15], Pearson and Wilks [11], Hsu [7]), for the equality of the means of k populations may be applied. For a single test of hypothesis (1′) we may use

$$W = \frac{|b'_{ij}|}{|b'_{ij} + m'_{ij} + a'_{ij}|} . \tag{11}$$

As before, if W is significant we may test whether the deviation is due to constant differences or to other types of difference among the scales by calculating

$$W_m = \frac{|b'_{ij}|}{|b'_{ij} + m'_{ij}|} , \tag{12}$$

and

$$W_t = \frac{|b'_{ij}|}{|b'_{ij} + a'_{ij}|} . \tag{13}$$

The flexibility of analysis of variance tests is not sacrificed, in particular we may test any desired subgroup of the treatments or of the scales. When there are only two scales the tests reduce to those given in section 3.

The tests are invariant under homogeneous linear transformations of the y's

which explains why the form of the subdivision of the scale comparisons is immaterial. In fact for purposes of computation it is not necessary to introduce the y's. By taking a simple transformation and expressing a'_{ij} in terms of a_{ij}, etc., we may express W directly in terms of the x's, as follows

$$W = \frac{\sum_{ij} B_{ij}}{\sum_{ij} (B + M + A)_{ij}}, \tag{14}$$

where B_{ij}, $(B + M + A)_{ij}$ are respectively the co-factors of the matrices (b_{ij}), $(b_{ij} + m_{ij} + a_{ij})$. Analogous expressions hold for W_m and W_t. In practice it will often be preferable to compute the y's in order that particular comparisons among the scale variates may be examined in detail.

The form of the frequency distribution has been worked out by Wilks [15]. For small values of n_1 and p, the test of significance can be referred to the recent tables of the significance levels of the incomplete Beta-function, Thompson [13], or to variance-ratio tables. Such cases are listed below, from Wilks [15] and Hsu [7]. In our notation, ν_1 is taken as $(n_1 + 1)$ in equation (11), as 1 in equation (12) and as n_1 in equation (13).

$\underline{p = 3, \nu_1 > 1}$: $f(W) \propto W^{\frac{1}{2}(n_2-3)}(1 - W^{\frac{1}{2}})^{\nu_1-1}$

$$: F\{2\nu_1, 2(n_2 - 1)\} = \frac{(n_2 - 1)(1 - W^{\frac{1}{2}})}{\nu_1 W^{\frac{1}{2}}},$$

$\underline{\nu_1 = 1}$: $f(W) \propto W^{\frac{1}{2}(n_2-p)}(1 - W)^{\frac{1}{2}(p-3)}$

$$: F\{p - 1, n_2 - p\} = \frac{(n_2 - p)(1 - W)}{(p - 1)W}.$$

This distribution applies to all tests made on the Mean, equation (12), and all cases where a single degree of freedom is isolated from the treatment comparisons.

$\underline{\nu_1 = 2}$: $f(W) \propto W^{\frac{1}{2}(n_2-p)}(1 - W^{\frac{1}{2}})^{p-2}$

$$: F\{2(p - 1), 2(n_2 - p + 2)\} = \frac{(n_2 - p + 2)(1 - W^{\frac{1}{2}})}{(p - 1)W^{\frac{1}{2}}}.$$

A tabulation of the distributions for four and five scales would be useful. Hsu [7] has shown that as n_2 becomes large, the distribution of $-n_2 \log W$ tends to that of χ^2 with $\nu_1(p - 1)$ degrees of freedom. In general, this approximation does not agree very well with the exact distributions above unless n_2 exceeds 60.

5. Interpretation as a problem in canonical correlations. As an introduction to the methods that will be used in testing the hypothesis of linearity, we may note that hypotheses (1′) and (2′) can be described in terms of canonical correlations. Fisher [5] has pointed out that the roots θ of the equation

$$| a_{ij} - \theta(a_{ij} + b_{ij}) | = 0, \tag{15}$$

are the squares of the *sample* canonical correlations between the x-variates and a set of n_1 dummy variates which represent the n_1 degrees of freedom among treatments. In order to obtain the corresponding equation for the *population* correlations, we may suppose that n_1 and p remain constant while the number of replicates r' and consequently n_2 increase without limit. After the removal of a common factor r', equation (15) becomes

$$|\psi_{ij} - \rho^2(\psi_{ij} + \nu\sigma_{ij})| = 0, \tag{16}$$

where

$$\psi_{ij} = \sum_{t=1}^{n_1+1} (\xi_{it} - \bar{\xi}_i)(\xi_{jt} - \bar{\xi}_j). \tag{17}$$

The value of the coefficient ν depends on the type of experimental design. For a randomized block layout, $\nu = n_1$ and for a simple group comparison $\nu = (n_1 + 1)$.

Now if hypothesis (2') is true, i.e., $\xi_{it} = \xi_i + \epsilon_t$, it follows that ψ_{ij} is independent of i and j. In this event equation (16) has $(p - 1)$ roots ρ^2 which are identically zero. The remaining root corresponds to the best discriminant function, Fisher [5], and does not vanish unless the treatments have no effects on any of the x-variates.

Let $\Sigma\beta_i x_i$ be a population canonical variate for the scale variables. The coefficients β_i satisfy the equations

$$\sum_j \beta_j\{\psi_{ij} - \rho^2(\psi_{ij} + \nu\sigma_{ij})\} = 0. \qquad i = 1, \cdots p. \tag{18}$$

For a zero root $\rho^2 = 0$ we have $\psi_{ij} =$ constant. Hence if a zero root is substituted, equation (18) degenerates into

$$\beta_1 + \beta_2 + \cdots + \beta_p = 0. \tag{19}$$

To summarize, hypothesis (2') specifies that (i) $(p - 1)$ of the population canonical correlations vanish and (ii) any variate $\Sigma\beta_i x_i$ is a canonical scale variate corresponding to a zero root, provided that equation (19) is satisfied. Analogous results hold for hypothesis (1'); in this case we replace the Treatments line of the analysis of variance by the (Treatments + Mean) line.

6. Test for linear relationship—two scales. We may assume $n_1 \geq 2$, otherwise no test of linearity is possible. If the values of α, β and γ in equations (3) are known, the problem can be reduced to that of testing hypothesis (1) or (2). Since this case is unlikely to be encountered frequently in practice, further details are omitted.

When α, β and γ are unknown, we may theoretically replace the variates x_1 and x_2 by $v_1 = \alpha x_1 + \beta x_2$ and $v_2 = \mu_1 x_1 + \mu_2 x_2$, where μ_1 and μ_2 are chosen so that v_1 and v_2 are independently distributed. If hypothesis (3) holds, it follows from (17) that in terms of the v's, $\psi_{11} = \psi_{12} = 0$. Since in addition $\sigma_{12} = 0$, the two roots of equation (16) are

$$\rho^2 = 0 \quad \text{and} \quad \rho^2 = \psi_{22}/(\psi_{22} + \nu\sigma_{22}). \tag{20}$$

Thus hypothesis (3) implies that one of the population canonical correlations vanishes. Unlike the previous case, however, we cannot construct the corresponding canonical variate, which requires knowledge of α and β.

The selection of a sample test criterion opens up some difficulties. Pending further elucidation of the problem, the natural choice seems to be the square r_2^2 of the lower sample canonical correlation, or the equivalent quantity $h_2 = r_2^2/(1 - r_2^2)$, where h_2 is the lower root of the equation:

$$|a_{ij} - hb_{ij}| = 0. \tag{21}$$

It appears likely, however, that r_2^2 and h_2 are not sufficient estimates of the corresponding population parameters.

When n_2 is large, Hsu [8] has shown that the distribution of $n_2 h_2$ tends to that of χ^2 with $(n_1 - 1)$ degrees of freedom. A considerable advance towards the small-sample distribution is obtainable from Madow [10], who developed an expression for the exact distribution of r_1^2 and r_2^2 when one of the population correlations is different from zero. In our notation this result, which is an important generalization of the distribution found by Fisher [5] and Girshick [6] may be written as follows:

$$\frac{(n_1 + n_2 - 2)!}{4\pi(n_1 - 2)!\,(n_2 - 2)!}\,(r_1^2 r_2^2)^{\frac{n_1-3}{2}}\,\{(1 - r_1^2)(1 - r_2^2)\}^{\frac{n_2-3}{2}}\,(r_1^2 - r_2^2)\,dr_1^2\,dr_2^2$$

$$\times (1 - \rho_1^2)^{\frac{1}{2}(n_1+n_2)} \int_{r_2^2}^{r_1^2} \frac{F\left(\frac{n_1 + n_2}{2}, \frac{n_1 + n_2}{2}, \frac{n_1}{2}, \rho_1^2 y\right) dy}{\sqrt{(r_1^2 - y)(y - r_2^2)}}, \tag{22}$$

where ρ_1 is the non-vanishing population correlation. It is evident from the form of (22) that the distribution of r_2^2 or h_2 involves ρ_1. The conditional distribution of h_2/h_1 may be relatively insensitive to changes in ρ_1, though even this distribution does not seem entirely independent of ρ_1.

When ρ_1 is unity, the small-sample distribution of h_2 is that of the ratio of two independent sums of squares, i.e., $h_2 = (n_1 - 1)e^{2z}/n_2$, with $(n_1 - 1)$ and n_2 degrees of freedom. This result is a particular case of a more general result proved in section 8. From (20) it is seen that ρ_1 is close to unity when ψ_{22} is large relative to σ_{22}, i.e., when the real differences among the treatments are large relative to the experimental errors. In the absence of a usable exact solution, the F-distribution may be a better approximation than the large-sample distribution of h_2 for data where r_1 is found to be close to unity, though proof of this statement is not yet available.

If it is desired to test hypothesis (3) with the additional assumption that $\gamma = 0$, we replace a_{ij} by $(a_{ij} + m_{ij})$ in equation (21) for h_2, and n_1 by $(n_1 + 1)$ in the distribution theory.

7. Connection with the method of least squares. The previous approach has an interesting connection with the method of least squares. We are required to test the linearity of relationship between $(n_1 + 1)$ pairs of means $(\bar{x}_{1t}, \bar{x}_{2t})$.

Both variates are subject to error and the errors are correlated; with r' replications the population variances and covariance of these means are σ_{11}/r', σ_{22}/r' and σ_{12}/r'. For these unknown quantities we have sample estimates b_{11}/n_2r', b_{22}/n_2r' and b_{12}/n_2r' respectively, derived from the Error line of the analysis of variance.

The procedure suggested by the method of least squares is to estimate the parameters of the line and use the deviations of the points $(\bar{x}_{1t}, \bar{x}_{2t})$ from the line for a test of linearity. If the population variances were known, the unknown quantities α, β, γ and ξ_{it} would be estimated by minimizing the quadratic form:

$$(23)\quad \sigma^{11}\sum_{t=1}^{n_1+1} r'(\bar{x}_{1t}-\xi_{1t})^2 + 2\sigma^{12}\sum_{t=1}^{n_1+1} r'(\bar{x}_{1t}-\xi_{1t})(\bar{x}_{2t}-\xi_{2t}) + \sigma^{22}\sum_{t=1}^{n_1+1} r'(\bar{x}_{2t}-\xi_{2t})^2,$$

subject to the linear relations (3). Here (σ^{ij}) is the matrix inverse to σ_{ij}. On substitution of the estimates, expression (23), which is positive definite, would serve as a "sum of squares" of deviations from the line and therefore as a test criterion. This criterion is of course a direct generalization of the weighted sum of squares which is used when the errors are independent.

Van Uven [14] gave an elegant method by which the sum of squares of deviations can be found directly, before solving for any of the unknown quantities. In our notation he showed that the sum of squares of deviations is the smaller root H_2 of the equation

$$(24)\qquad |\, a_{ij} - H\sigma_{ij} \,| = 0,$$

where a_{ij} is as before the treatments sum of squares or products.

Suppose that in default of knowledge of the σ_{ij} we derive the weights from the sample estimates b_{ij}/n_2; i.e., we minimize (23) with b^{ij} in place of σ^{ij}, where $(b^{ij}) = (b_{ij}/n_2)^{-1}$. In this case the method of Van Uven shows that the sum of squares of deviations from the best-fitting line is the smaller root H_2' of the equation

$$(25)\qquad \left| a_{ij} - \frac{H'}{n_2} b_{ij} \right| = 0.$$

Comparing (25) with (21) we find $H_2' = n_2 l_2$. Consequently the least squares approach, with sample weights substituted in (23) for the unknown true weights, leads to l_2 as a test criterion. Further, Hsu's [8] proof that the distribution of $n_2 l_2$ tends to χ^2 with $(n_1 - 1)$ degrees of freedom establishes for this case the standard least-squares result for the distribution of the residual sum of squares: —namely that when the population weights are known, the residual sum of squares is distributed as χ^2, with degrees of freedom equal to the number of points, $2(n_1 + 1)$, *minus* the number of independent unknowns, $(n_1 + 3)$. By a transformation of the x-variates to independent variables, this result can be obtained alternatively from a theorem by Deming [2].

8. Test for linear relationship—more than two scales. The extension of hypothesis (3) to the case of p scales can be expressed by means of the equations

$$(3') \qquad \alpha_i\xi_{1t} + \beta_i\xi_{it} = \gamma_i : \qquad (i = 2, \cdots p)(t = 1, \cdots n_1 + 1).$$

The equations, $(p - 1)(n_1 + 1)$ in number, postulate a linear relation between x_1 and every other variate and consequently imply a linear relation between any pair of variates x_i and x_j

Consider the variates $v_i = \alpha_i x_{1t} + \beta_i x_{it}$, $(i = 2, \cdots p)$ For v_1 we choose the linear function of the x's which is independent of $v_2, \cdots v_p$. Thus in equation (16) for the population canonical correlations we have $\psi_{ij} = 0$, $(i, j, \geq 2)$ and $\sigma_{1j} = 0$, $(j > 1)$. It follows that all roots of equation (16) are zero except one, the non-vanishing root being $\rho^2 = \psi_{11}/(\psi_{11} + \nu\sigma_{11})$. If each treatment denotes a separate population, hypothesis (3′) is therefore identical with Fisher's hypothesis [4], that the populations are *collinear*.

As a test criterion for this hypothesis Fisher has suggested the sum of the roots of equation (21), excluding the highest root, i e, $V' = \Sigma h_i = \Sigma r_i^2/(1 - r_i^2)$. If $n_1 \geq p$ the sum extends over $(p - 1)$ roots, while if $n_1 < p$ the sum extends over $(n_1 - 1)$ roots. For computational purposes it may be more expeditious to form this sum by subtraction Hsu [7] has pointed out that the sum of all roots is given by $V = \sum_{ij} b^{ij}a_{ij}$, which is obtained readily when the inverse of (b_{ij}) has been calculated The largest root of (21) is then found and subtracted from V.

Fisher [4] also suggested that when equations (3′) hold, the distribution of V' is approximately that of χ^2 with $(p - 1)(n_1 - 1)$ degrees of freedom. This result has been confirmed by Hsu [8] as the limiting form of the V' distribution when n_2 tends to infinity. As in the case of two scales, the small-sample distribution is as yet unknown; it presumably contains ρ_1, the non-vanishing correlation, as a nuisance parameter.

Some progress towards the small-sample distribution can be made without difficulty in the case where $\rho_1 = 1$. For then v_1 must have a zero Error sum of squares in every sample from the population, i e, v_1 is constant within any given treatment. Consequently (i) $b_{1i} = 0$ for $i = 1, \cdot \; p$, and (ii) a_{1j}^2/a_{11} is a single degree of freedom from the Treatments sum of squares of v_j. On account of conditions (i), equation (21) reduces to

$$(26) \qquad \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{12} & a_{22} - hb_{22} & \cdot \; \cdot & a_{2p} - hb_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ a_{1p} & a_{2p} - hb_{2p} & \cdots & a_{pp} - hb_{pp} \end{vmatrix} = 0.$$

Subtract a_{1i}/a_{11} times the first row from the ith row, for $i = 2, \cdots p$. We see that one root is infinite; the rest are the roots of the equation

$$(27) \qquad |a''_{ij} - hb_{ij}| = 0, \qquad i, j = 2, \cdots p,$$

where $a''_{ij} = a_{ij} - a_{1i}a_{1j}/a_{11}$

If hypothesis (3′) holds, the quantities a''_{ij} follow the Wishart distribution [16] with $(n_1 - 1)$ degrees of freedom. Hence the joint distribution of $h_2, \cdots h_p$ or h_{n_1}, is that which is obtained when all the population canonical correlations vanish, with $(n_1 - 1)$ in place of n_1. For $n_1 \geq p$, the distribution function (apart from the constant term) is:

$$\prod_{i=2}^{p}\left[h_i^{\frac{1}{2}(n_1-p-1)}(1+h_i)^{-\frac{1}{2}(n_1+n_2-1)}\left\{\prod_{j=i+1}^{p}(h_i-h_j)\right\}\right]. \tag{28}$$

For two scales, $(p = 2)$, we reach the result mentioned in section 6, that $V' = h^2$ is distributed as $(n_1 - 1)e^{2\theta}/n_2$. This result can also be obtained directly from (27). When $p = 3$, the distribution of V' is obtainable from a result by Hsu [7].

9. Measures of relative sensitivity. We propose to discuss briefly the estimation of the relative sensitivity of two scales and to indicate the types of distribution that are involved. If there are only two treatments, t, t', an appropriate definition of the true sensitivity of the ith scale is

$$\frac{(\xi_{it'} - \xi_{it})^2}{2\sigma_{ii}}, \tag{29}$$

or some simple function of this quantity. In justification, we may observe that for a fixed number of replicates, the power function of the t-test in the ith scale depends entirely on this quantity. An unbiased sample estimate is

$$\frac{(n_2 - 2)(\bar{x}_{it'} - \bar{x}_{it})^2}{2b_{ii}} - \frac{1}{r'}, \tag{30}$$

where r' is the number of replicates. Since (30) involves a non-central variance ratio, confidence limits for the true sensitivity can be found from Fisher's Type C distribution, Fisher [3].

It follows from (3) and (29) that if two scales are linearly related (including the case of equivalence) their relative sensitivity is constant for all treatment comparisons. For scale 1 relative to scale 2 the sensitivity is measured by $\beta^2\sigma_{22}/\alpha^2\sigma_{11}$.

If the scales are equivalent, apart possibly from a constant difference, this quantity reduces to $\varphi = \sigma_{22}/\sigma_{11}$, for which $F = b_{22}/b_{11}$ serves as a sample estimate. A test of significance of the sample ratio and confidence limits for the true ratio may be obtained from Pitman [12], who showed that

$$\left(\frac{F}{\varphi} - 1\right) \Big/ \sqrt{\left(\frac{F}{\varphi} + 1\right)^2 - \frac{4r_{12}^2 F}{\varphi}}, \tag{31}$$

follows the distribution of a sample correlation coefficient from $(n_2 + 1)$ pairs of observations. In (31), $r_{12}^2 = b_{12}^2/b_{11}b_{22}$. The same procedure may be used whenever α and β are known.

When α and β are unknown, a sample estimate of the relative sensitivity is b^2b_{22}/a^2b_{11}, where $(ax_1 + bx_2)$ is the discriminant function which corresponds to

the lower root of equation (21). We have not been able to reach the distribution of this estimate. Confidence limits for the relative sensitivity can, however, be obtained when n_2 is sufficiently large so that σ_{11} and σ_{22} may be assumed known. For in that case the problem reduces to that of finding confidence limits for β^2/α^2. Now if α, β are the true coefficients, the quantity

$$\frac{\alpha^2 a_{11} + 2\alpha\beta a_{12} + \beta^2 a_{22}}{\alpha^2 b_{11} + 2\alpha\beta b_{12} + \beta^2 b_{22}}, \tag{32}$$

follows the $n_1 e^{2z}/n_2$ distribution. Any proposed values of α and β which make (32) significant are rejected by the evidence of the sample. By equating (32) to the desired significance level of $n_1 e^{2z}/n_2$, we get a quadratic equation for the two limits of β/α. The limits will not be narrow unless the treatment effects are large.

If the relation between the scales is non-linear, and the assumption of a constant error variance throughout an individual scale is valid, the relative sensitivity differs for different treatment comparisons. Even in this event estimates of relative sensitivity may be of interest. Attention might be restricted to a single degree of freedom from the treatment comparisons, in which case the definition for two treatments could be applied.

Alternatively an estimate might be wanted of the *average* relative sensitivity over all treatment comparisons. For a given number of replicates, the power function of the variance-ratio test of the treatment effects in the ith scale depends only on the quantity

$$\frac{\sum_t (\xi_{it} - \bar{\xi}_i)^2}{\sigma_{ii}}. \tag{33}$$

Consequently this quantity, which is an extension of (29), might be chosen as a measure of average sensitivity. The corresponding generalization of the unbiased sample estimate (20) is

$$\frac{(n_2 - 2)a_{ii}}{n_1 r' b_{ii}} - \frac{1}{r'}. \tag{34}$$

Since the quantity a_{ii}/b_{ii} is a multiple of a non-central variance ratio, the comparison of two scales involves a test of significance of the hypothesis that two non-central variance ratios are equal.

10. Summary. This paper discusses the analysis of data obtained when the results of a replicated experiment are measured on several different scales which we wish to compare. Recent work in multivariate analysis provides tests of the hypothesis that the treatment effects are the same in all scales, and of the hypothesis that the scales are linearly related. When the number of Error degrees of freedom is large, the significance levels of these tests are obtainable from the standard tables. For small sample tests, further investigation and

tabulation of certain distributions will be needed, particularly that of the sample canonical correlations when one population correlation differs from zero.

A brief discussion is given of methods for comparing the relative sensitivity of two scales.

REFERENCES

[1] K. H. Coward, *The Biological Standardization of the Vitamins*, 1937.

[2] W E. Deming, "The chi-test and curve-fitting," *Jour Amer. Stat. Assn.*, Vol. 29 (1934), pp. 372–382

[3] R A Fisher, "The general sampling distribution of the multiple correlation coefficient," *Proc. Roy. Soc. A*, Vol 121 (1928), pp. 654–673.

[4] R. A. Fisher, "The statistical utilization of multiple measurements," *Annals of Eugenics*, Vol. 8 (1938), pp 376–386

[5] R. A Fisher, "The sampling distribution of some statistics obtained from non-linear equations," *Annals of Eugenics*, Vol 9 (1939), pp 238–249

[6] M A Girshick, "On the sampling theory of the roots of determinental equations," *Annals of Math. Stat*, Vol. 10 (1939), pp 203–224

[7] P L. Hsu, "On generalized analysis of variance (I)," *Biometrika*, Vol 31 (1940), pp. 221–237.

[8] P. L Hsu, "The problem of rank and the limiting distribution of Fisher's test function," *Annals of Eugenics*, Vol. 11 (1941), pp. 39–41

[9] D N. Lawley, "A generalization of Fisher's z-test," *Biometrika*, Vol. 30 (1938), pp. 180–187.

[10] W G Madow, "Contributions to the theory of multivariate statistical analysis," *Trans. Amer Math. Soc.*, Vol 44 (1938), p 490

[11] E S. Pearson and S. S. Wilks, "Methods of statistical analysis appropriate for k samples of two variables," *Biometrika*, Vol. 25, (1933), pp 353–378

[12] E J G. Pitman, "A note on normal correlation," *Biometrika*, Vol. 31 (1939), pp 9–12

[13] C. M. Thompson, "Tables of percentage points of the incomplete Beta-function," *Biometrika*, Vol 32 (1942), pp. 151–181

[14] M. J. Van Uven, "Adjustment of N points (in n-dimensional space) to the best linear $(n-1)$ dimensional space," *Proc. Koninklizke Akad van Wetenschappen te Amsterdam*, Vol 33 (1930), pp. 143–157.

[15] S. S Wilks, "Certain generalizations in the analysis of variance," *Biometrika*, Vol. 24 (1932), pp 471–494

[16] J. Wishart, "The generalized product moment distribution in samples from a normal multivariate population," *Biometrika*, Vol 20A (1928), pp. 32–52

ON STOCHASTIC LIMIT AND ORDER RELATIONSHIPS

By H. B. Mann and A. Wald[1]

Columbia University

1. Introduction. The concept of a stochastic limit is frequently used in statistical literature. Writers of papers on problems in statistics and probability usually prove only those special cases of more general theorems which are necessary for the solution of their particular problems. Thus readers of statistical papers are confronted with the necessity of laboriously ploughing through details, a task which is made more difficult by the fact that no uniform notation has as yet been introduced. It is therefore the purpose of the present paper to outline a systematic theory of stochastic limit and order relationships and at the same time to propose a convenient notation analogous to the notation of ordinary limit and order relationships. The theorems derived in this paper are of a more general nature and seem to contain to the authors' knowledge all previous results in the literature. For instance the so-called δ-method for the derivation of asymptotic standard deviations and limit distributions, also two lemmas by J. L. Doob [1] on products, sums and quotients of random variables and a theorem derived by W. G. Madow [2] are special cases of our results. It is hoped that such a general theory together with a convenient notation will considerably facilitate the derivation of theorems concerning stochastic limits and limit distributions. In section 2 we define the notion of convergence in probability and that of stochastic order and derive 5 theorems of a very general nature. Section 2 contains 2 corollaries of these general theorems which have so far been most important in applications.

We shall frequently need the concept of a vector. A vector $a = (a^1, \cdots, a^r)$ is an ordered set of r numbers $a^1, \cdots, a^r$. The numbers $a^1, \cdots, a^r$ are called the components of a. If the components are random variables then the vector is called a random vector.

We shall generally denote by a, b constant vectors by x, y random vectors and by $a^1, \cdots, a^r, x^1, \cdots, x^r$ their components. Differing from the usual practice we shall put $|a| = (|a^1|, \cdots, |a^r|)$ and we shall write $a < b$ or $a \leq b$ if $a^i < b^i$ or $a^i \leq b^i$ for every i. This notation saves a great amount of writing, since all our theorems except theorem 4 are valid for sequences of any number of jointly distributed variates.

We shall review here the ordinary order notation. In all that follows let $f(N)$ be a positive function defined for all positive integers N.

---

[1] Research under a grant-in-aid from the Carnegie Corporation of New York

We write

$a_N = o[f(N)]$ if $\lim_{N\to\infty} a_N/f(N) = 0$.

$a_N = O[f(N)]$ if $|a_N| \leq Mf(N)$ for all N and a fixed $M > 0$.

$a_N = \Omega[f(N)]$ if $0 < M'f(N) \leq |a_N| \leq Mf(N)$ for almost all N and for two fixed numbers $M > M' > 0$.

$a_N = \omega[f(N)]$ if $0 < Mf(N) \leq |a_N|$ for almost all N and a fixed $M > 0$.

For instance, $\log N = o(N^\epsilon)$ for every $\epsilon > 0$, or $\sin N/N = O(1/N)$, $3 + 4\cdot N/(4 + 8\sqrt{N}) = \Omega(\sqrt{N})$ $5/\sin N = \omega(1)$.

For any statement V we shall denote by $P(V)$ the probability that V holds.

2. General theorems on stochastic limit and order relationships.

DEFINITION 1. *We write* $\operatorname{plim}_{N\to\infty} x_N = 0$. *(In words x_N converges in probability to 0 with increasing N) if for every $\epsilon > 0$* $\lim P(|x_N| \leq \epsilon) = 1$. *Further* $\operatorname{plim}_{N\to\infty} x_N = x$ *if* $\operatorname{plim}_{N\to\infty} (x_N - x) = 0$.

DEFINITION 2. *We write $x_N = o_p[f(N)]$ (x_N is of probability order $o[f(N)]$) if* $\operatorname{plim}_{N\to\infty} x_N/f(N) = 0$.

DEFINITION 3. *We write $x_N = O_p[f(N)]$ (x_N is of probability order $O[f(N)]$) if for each $\epsilon > 0$ there exists an $A_\epsilon > 0$ such that $P(|x_N| \leq A_\epsilon f(N)) \geq 1 - \epsilon$ for all values of N.*

DEFINITION 4 *$x_N = \Omega_p[f(N)]$ if for each $\epsilon > 0$ there exist two numbers $A_\epsilon > 0$ and $B_\epsilon > 0$ and an integer N_ϵ such that $P[A_\epsilon f(N) \leq |x_N| \leq B_\epsilon f(N)] \geq 1 - \epsilon$ for all $N \geq N_\epsilon$.*

DEFINITION 5. *$x_N = \omega_p[f(N)]$ if for every $\epsilon > 0$ there exists an $A_\epsilon > 0$ and an integer N_ϵ such that $P[A_\epsilon f(N) < |x_N|] \geq 1 - \epsilon$ for all $N \geq N_\epsilon$.*

Let E denote a vector space. For any subset E' of E the symbol $a \subset E'$ will mean that a is an element of E'.

Since $P(x \subset E_1 \,\&\, x \subset E_2) \geq P(x \subset E_1) - P(x \not\subset E_2)$ we evidently have

LEMMA 1. *If $P(x \subset E_1) \geq 1 - \epsilon$, $P(x \subset E_2) \geq 1 - \epsilon'$, then $P(x \subset E_1;\ x \subset E_2) \geq 1 - \epsilon - \epsilon'$.*

We now put $O^1 = o$, $O^2 = O$, $O^3 = \Omega$, $O^4 = \omega$.

THEOREM 1. *For every $\epsilon > 0$ let $\{R_N(\epsilon)\}$ be a sequence of subsets of the r-dimensional Cartesian space such that $P(x_N \subset R_N(\epsilon)) \geq 1 - \epsilon$ for all N greater than a certain integer N_ϵ. Let $\{g_N(x)\}$ be a sequence of functions of $x = (x^1, x^2, \cdots x^r)$ such that $g_N(a_N) = O^i[f(N)]$ for any $\epsilon > 0$ and for any sequence $\{a_N\}$ for which $a_N \subset R_N(\epsilon)$. Then we have $g_N(x_N) = O_p^i[f(N)]$.*

PROOF: For $i = 1, 2, 3$, there exists a positive integer $\bar{N}_\epsilon$ such that $|g_N(a)|$ is a bounded function of a in $R_N(\epsilon)$ for $N > \bar{N}_\epsilon$. For otherwise we could construct a sequence $\{a_N\}$ with a_N in $R_N(\epsilon)$ such that $|g_N(a_N)| > Mf(N)$ for any M and for infinitely many values of N which contradicts the hypothesis of our theorem. Hence there exists an $\bar{N}_\epsilon$ such that for $N > \bar{N}_\epsilon$ the function $|g_N(a)|$ is bounded in $R_N(\epsilon)$. Let $M_N(\epsilon)$ be the l.u.b. of $|g_N(a)|/f(N)$ in $R_N(\epsilon)$. We can construct a sequence $\{a_N\}$ with $a_N \subset R_N(\epsilon)$ such that $|g_N(a_N)|/f(N) \geq M_N(\epsilon)/2$ for all $N > \bar{N}_\epsilon$. Hence for $i = 2, 3$ the sequence $M_N(\epsilon)$ must be bounded and for

$i = 1$ we must have $\lim_{N\to\infty} M_N(\epsilon) = 0$. Let $M(\epsilon)$ be the l.u.b. of $M_N(\epsilon)$. For $i = 3, 4$ one shows in exactly the same manner the existence of a g.l.b. $\bar{M}(\epsilon)$ of $|g_N(a)|/f(N)$ if $a \subset R_N(\epsilon)$ and for $N > N'_\epsilon$. Hence for sufficiently large N we have

$$P[|g_N(x_N)| \leq M_N(\epsilon)f(N)] \geq 1 - \epsilon \text{ with } \lim_{N\to\infty} M_N(\epsilon) = 0 \text{ for } i = 1,$$

$$P[|g_N(x_N)| \leq M(\epsilon)f(N)] > 1 - \epsilon \qquad \text{for } i = 2,$$

$$P[\bar{M}(\epsilon)f(N) \leq |g_N(x_N)| \leq M(\epsilon)f(N)] \geq 1 - \epsilon \qquad \text{for } i = 3,$$

$$P[\bar{M}(\epsilon)f(N) \leq |g_N(x_N)|] \geq 1 - \epsilon \qquad \text{for } i = 4.$$

For $i = 2$ the existence of an $M'(\epsilon)$ such that $P[|g_N(x_N)| \leq M'(\epsilon)f(N)] \geq 1 - \epsilon$ for all N follows easily from this result. Hence our theorem is proved.

COROLLARY 1. *If $x'_N = O'_p[f_j(N)]$ for $j = 1, 2, \cdots, r$ and $\{\bar{R}_N(\epsilon)\}$ is a sequence of subsets of the k-dimensional space $y^1, y^2, \cdots y^k$ such that $P[y_N \subset \bar{R}_N(\epsilon)] \geq 1 - \epsilon$ for sufficiently large N, and if $\{g_N(x^1, x^2, \cdots x^r, y^1, y^2, \cdots y^k,)\}$ is a sequence of functions of $x^1, x^2, \cdots x^r, y^1, y^2, \cdots y^k$ such that for any $\epsilon > 0$ we have $g_N(a_N, b_N) = O'[f(N)]$ for every sequence $\{a_N, b_N\}$ with $a'_N = O'[f_j(N)]$ $(j = 1, 2, \cdots, r)$ and $b_N \subset \bar{R}_N(\epsilon)$, then $g_N(x_N, y_N) = O'_p[f(N)]$.*

PROOF: It follows from Lemma 1, the definition of the relation $x'_N = O'_p[f_j(N)]$ and the hypothesis of our corollary that for any $\epsilon > 0$ there exists a sequence of subsets $\{R_N(\epsilon)\}$ of the space $x^1, \cdots, x^r, y^1, \cdots, y^k$ which satisfies the conditions of Theorem 1 with respect to the sequence of functions $\{g_N\}$. Hence Corollary 1 is an immediate consequence of Theorem 1.

Corollary 1 implies *inter alia* that all operational rules for the ordinary order and limit relations are also applicable to stochastic limit and order relations. For instance $o[f(N)]/\Omega[g(N)] = o[f(N)/g(N)]$. Hence also $o_p[f(N)]/\Omega_p[g(N)] = o_p[f(N)/g(N)]$

DEFINITION 6. *For any N let R_N be a region, $f_N(a)$ a function defined on R_N. The sequence $\{f_N(a)\}$ will be said to be uniformly continuous with respect to $\{R_N\}$ if the following condition is fulfilled. For every $\epsilon > 0$ there exists a vector $\delta > 0$ such that for almost all N*

$$|f_N(a + \bar{\delta}) - f_N(a)| \leq \epsilon \qquad \text{for any } |\bar{\delta}| < \delta, \text{ and for any } a \subset R_N$$

THEOREM 2. *Let $\operatorname{plim}_{N\to\infty} (x_N - y_N) = 0$. For every $\epsilon > 0$ let $\{R_N(\epsilon)\}$ be a sequence of subsets of the r-dimensional vector space such that for almost all N we have $P[y_N \subset R_N(\epsilon)] \geq 1 - \epsilon$. If the sequence of functions $\{f_N(a)\}$ is uniformly continuous with respect to $\{R_N(\epsilon)\}$ for every $\epsilon > 0$, then $\operatorname{plim}_{N\to\infty} [f_N(x_N) - f_N(y_N)] = 0$.*

PROOF: We have $f_N(x_N) - f_N(y_N) = f_N(y_N + z_N) - f_N(y_N)$ where $z'_N = o(1)$ *for* $j = 1, \cdots, r$. Because of the uniform continuity of $f_N(a)$ with respect to $R_N(\epsilon)$ we see that for every sequence $\{a_N, b_N\}$ with $a_N \subset R_N(\epsilon)$ and $b'_N = o(1)$ $(j = 1, 2, \cdots, r)$.

$$f_N(a_N + b_N) - f_N(a_N) = o(1).$$

Hence Theorem 2 follows from Corollary 1.

In the following we shall abbreviate "cumulative distribution function" by d.f.

DEFINITION 7. *Let $\{x_N\}$ be a sequence of random variables. Let F_N be the d.f. of x_N. Let x have the distribution F. We shall write $d_\infty(x_N) = d(x)$ if $\lim_{N\to\infty} F_N = F$ in every continuity point of F.*

THEOREM 3 *Let $\operatorname{plim}_{N\to\infty}(x_N - y_N) = 0$ and $d_\infty(y_N) = d(y)$, then $d_\infty(x_N) = d(y)$.*

PROOF: Let G_N, F_N be the d.f.'s of x_N, y_N resp. For any $\delta > 0$ we have

$$P(y_N \leq a + \delta) \geq P(x_N \leq a; y_N \leq a + \delta) \geq P(x_N \leq a; |y_N - x_N| \leq \delta)$$
$$\geq P(x_N \leq a) - P(|y_N - x_N| > \delta),$$
$$P(x_N \leq a) \geq P(x_N \leq a; y_N \leq a - \delta) \geq P(y_N \leq a - \delta)$$
$$- P(|x_N - y_N| > \delta).$$

Hence since $P(y_N \leq a) = F_N(a)$, $P(x_N \leq a) = G_N(a)$, $\lim_{N\to\infty} P(|x_N - y_N| > \delta) = 0$ we have lim. sup. $F_N(a + \delta) \geq$ lim. sup $G_N(a) \geq$ lim. inf. $G_N(a) \geq$ lim. inf. $F_N(a - \delta)$.

If $a + \delta$ and $a - \delta$ are continuity points of F we have

$$F(a + \delta) \geq \text{lim. sup } G_N(a) \geq \text{lim. inf. } G_N(a) \geq F(a - \delta).$$

For any $\delta_0 > 0$ there exists a positive $\delta < \delta_0$ such that $a - \delta$ and $a + \delta$ are continuity points of F. Hence we can choose δ arbitrarily small and if a is a continuity point of F we must have

$$\text{lim. } G_N(a) = F(a).$$

THEOREM 4. *Let x_N, y_N be two sequences of one-dimensional vectors and let $\operatorname{plim}_{N\to\infty}(x_N - y_N) = 0$. Let F_N, G_N be the cumulative distribution functions of x_N and y_N respectively. Let $R_N(\epsilon)$ be the set of points a for which $|F_N(a) - G_N(a)| > \epsilon$. Let $M_N(\epsilon)$ be the Lebesgue measure of this set. Then $\lim_{N\to\infty} M_N(\epsilon) = 0$ for every $\epsilon > 0$.*

We first prove the following lemma

LEMMA 2. *Let δ, ϵ be any arbitrary positive numbers and let f be a distribution function. The set of points a for which $f(a + \delta) - f(a) \geq \epsilon$ has at most the Lebesgue measure δ/ϵ*

PROOF: The points a for which $f(a + \delta) - f(a) \geq \epsilon$ must have a lower bound $\bar{a}$. Otherwise we could find infinitely many such points whose distance from each other is more than δ. But this contradicts the requirement that $f(\infty) = 1$. Let a_1 be the g.l.b. of the a's. Then for any $\eta > 0$ in the interval $(a_1 \leq x \leq a_1 + \delta + \eta)$ the value of F increases at least by the amount ϵ Let now a_2 be the g l.b. of the a's outside of this interval We continue our construction by constructing the interval $(a_2 \leq x \leq a_2 + \delta + \eta)$ and so forth But after at most

$1/\epsilon$ such steps the construction must stop. Hence all points a for which $f(a+\delta) - f(a) \geq \epsilon$ are contained in at most $1/\epsilon$ intervals of length $\delta + \eta$. Hence since η was arbitrary the Lebesgue measure of this set is at most δ/ϵ.

We come now to the proof of our theorem. We have

$$P(x_N \leq a) \geq P(x_N \leq a;\, y_N \leq a+\delta) \geq P(x_N \leq a) - P(|x_N - y_N| > \delta),$$

$$P(y_N \leq a+\delta) \geq P(x_N \leq a;\, y_N \leq a+\delta) \geq P(y_N \leq a+\delta) - P(|x_N - y_N| > \delta) - P(a \leq x_N \leq a+2\delta).$$

Therefore

$$P(x_N \leq a,\, y_N \leq a+\delta) = P(x_N \leq a) - \bar{\theta}_N P(|x_N - y_N| > \delta)$$

$$= P(y_N \leq a+\delta) - \bar{\theta}'_N P(|x_N - y_N| > \delta) - \bar{\theta}'_N P(a \leq x_N \leq a+2\delta),$$

where $0 \leq \bar{\theta}_N \leq 1$, $0 \leq \bar{\theta}'_N \leq 1$. Hence

$$P(y_N \leq a+\delta) = P(x_N \leq a) + \theta_N P(|x_N - y_N| > \delta) + \theta'_N[F_N(a+2\delta) - F_N(a)]$$

where $|\theta_N|, |\theta'_N| \leq 1$.

By hypothesis we have $P(|x_N - y_N| \geq 1/m) \leq 1/m$ for almost all N and every integer m Hence we can choose a sequence $\{\delta_N\}$ with $\delta_N > 0$ in such a way that $\lim_{N\to\infty} \delta_N = 0$, $\lim_{N\to\infty} P(|x_N - y_N| \geq \delta_N) = 0$. We can then choose N_ϵ so that $P(|x_N - y_N| \geq \delta_N) \leq \epsilon/3$ for $N \geq N_\epsilon$ Applying Lemma 2 we see that except for a set of measure at most $6\,\delta_N/\epsilon$ we have $F_N(a+2\delta_N) - F_N(a) \leq \epsilon/3$. Similarly the set of points for which $g_N(a+\delta_N) - g_N(a) \geq \epsilon/3$ has at most the Lebesgue measure $3\,\delta_N/\epsilon$ Hence, except in a set of points whose measure is at most $9\,\delta_N/\epsilon$, we have

$$|G_N(a) - F_N(a)| \leq \epsilon,$$

and this completes the proof of Theorem 4

THEOREM 4a. *Let* $\operatorname{plim}_{N\to\infty}(x_N - y_N) = 0$. *Let* F_N, G_N *be the distribution functions of* x_N, y_N *respectively. Furthermore, let* $R_N(\epsilon)$ *be the set of points inside an* r*-dimensional cube where* $|F_N - G_N| \geq \epsilon$ *and let* $M_N(\epsilon)$ *be the Lebesgue measure of* $R_N(\epsilon)$, *then* $\lim_{N\to\infty} M_N(\epsilon) = 0$.

We prove first

LEMMA 2a. *Let* $\delta = (\delta^1, \delta^2, \cdots, \delta^r) > 0$ *and max.* $\delta^i = d$. *Let* I *be the cube defined by* $(-A \leq x^i \leq A,\ i = 1, 2, \cdots r)$. *Let furthermore* f *be a d.f. Then the Lebesgue measure of the points* a *in* I *for which* $f(a+\delta) - f(a) \geq \epsilon$ *is at most* $dr^2 A^{r-1}/\epsilon$.

PROOF: Let $f_1(x^1), f_2(x^2), \cdots f_r(x^r)$ be the marginal distributions of $x^1, x^2, \cdots x^r$ respectively. It follows from Lemma 2 that the linear Lebesgue measure of those numbers a^i for which $f_i(a^i + \delta^i) - f_i(a^i) \geq \epsilon/r$ is smaller than rd/ϵ. We form the set $(x^i = a^i \;\&\; x \subset I)$ for every such a^i and for $i = 1, 2, \cdots r$. The

Lebesgue measure of the sum $R(\epsilon)$ of all these sets is at most $r^2 dA^{r-1}/\epsilon$. We shall show that $R(\epsilon)$ contains all points a inside I for which $f(a+\delta) - f(a) \geq \epsilon$. We have

$$f(a^1 + \delta^1, a^2 + \delta^2, \cdots, a^r + \delta^r) - f(a^1, a^2, \cdots a^r) = \Delta_1 + \Delta_2 + \cdots + \Delta_r,$$

where $\Delta_i = f(a^1, a^2, \cdots a^{i-1}, a^i + \delta^i, \cdots a^r + \delta^r) - f(a^1, \cdots a^i, a^{i+1} + \delta^{i+1}, \cdots a^r + \delta^r)$. If $f(a+\delta) - f(a) \geq \epsilon$ then we must have for at least one i

$$\Delta_i \geq \epsilon/r.$$

But Δ_i is the probability of a subset of the set $T = (a^i \leq x^i \leq a^i + \delta^i)$ and $f_i(a^i + \delta^i) - f_i(a^i)$ is the probability of T itself. Hence

$$\epsilon/r \leq \Delta_i \leq f_i(a^i + \delta^i) - f_i(a^i),$$

and if $(a^1, a^2, \cdots a^r)$ is in I then it is contained in $R(\epsilon)$. Hence Lemma 2a is proved.

The proof of Theorem 4a using Lemma 2a is similar to that of Theorem 4 and therefore it is omitted.

The Jordan measure of a set R with respect to the distribution function F is defined as follows. We consider only intervals whose boundary points are continuity points of F. We cover R with the sum I of a finite number of intervals. (The intervals themselves may also be infinite. For instance the sets $a \leq x < \infty$, $a < x < \infty$ are also considered intervals.) We consider $M(I) = \int_I dF$ for every I covering R. The g.l.b of all such $M(I)$ is called the exterior Jordan measure $M(R)$ of R. Similarly we consider all sums $\bar{I}$ of a finite number of intervals which are contained in R. The l.u.b. of $\int_{\bar{I}} dF$ is called the interior Jordan measure $\bar{M}(R)$ of R. If $M(R) = \bar{M}(R)$ then $M(R)$ is called the Jordan measure of R.

LEMMA 3. *Let $F_N(x)$ be a sequence of d.f.'s such that $\lim_{N\to\infty} F_N(x) = F(x)$ in every continuity point of $F(x)$. Let $h(x)$ be a bounded function such that the discontinuity points of $h(x)$ have the Jordan measure 0 with respect to F and such that $\int_{-\infty}^{+\infty} h(x)\, dF_N(x)$ and $\int_{-\infty}^{+\infty} h(x)\, dF(x)$ exist. Then $\lim_{N\to\infty} \int_{-\infty}^{+\infty} h(x)\, dF_N(x) = \int_{-\infty}^{+\infty} h(x)\, dF(x)$.*

PROOF: There is only an enumerable set of hyperplanes parallel to the plane $x^i = 0$ which have positive probability with respect to F. Hence we can find for every δ an interval net whose cells have a diameter at most δ and such that the boundary points of every cell are continuity points of F.

We first determine a closed finite interval I such that $\int_I dF(x) \geq 1 - \frac{\epsilon}{2}$ and such that the boundary points of I are continuity points of F. We further determine a sum I' of a finite number of open intervals such that I' contains all discontinuity points of h, $\int_{I'} dF(x) \leq \frac{\epsilon}{2}$, and such that the boundary of I' does

not contain any discontinuity points of F. All this is possible by hypothesis and because the set of hyperplanes with positive probability is enumerable. Let R be the subset of I consisting of all points of I which are not contained in I'. R is a closed set and can be decomposed into a finite number of intervals. The function h is continuous in R and therefore uniformly continuous. We can therefore cover R by a finite set of intervals such that the variation of h in every interval is less than ϵ and such that the boundary points of each interval are continuity points of F. Let $I_1, I_2, \cdots I_k$ be such a finite set of intervals. Let x_j be any point in I_j. We have

$$
\begin{aligned}
|H_N| &= \left| \int_{-\infty}^{+\infty} h(x)\, dF_N(x) - \int_{-\infty}^{+\infty} h(x)\, dF(x) \right| = \left| \sum_{j=1}^{k} \int_{I_j} [h(x) - h(x_j)]\, dF_N(x) \right. \\
&\quad - \sum_{j=1}^{k} \int_{I_j} [h(x) - h(x_j)]\, dF(x) + \sum_{j=1}^{k} h(x_j) \left[\int_{I_j} dF_N(x) - \int_{I_j} dF(x) \right] \\
&\quad \left. + \int_{x \notin R} h(x)\, dF_N(x) - \int_{x \notin R} h(x)\, dF(x) \right| \\
&\leq \epsilon + \epsilon + \sum_{j=1}^{k} h(x_j) \left[\int_{I_j} dF_N(x) - \int_{I_j} dF(x) \right] \\
&\quad + \text{max. } h(x) \left[\int_{x \notin R} dF_N(x) + \epsilon \right].
\end{aligned}
$$

But $\lim_{N \to \infty} \int_R dF_N(x) \geq 1 - \epsilon$. Hence

$$\text{lim. sup. } H_N \leq 2\epsilon + 2\epsilon \text{ max. } h(x) .$$

Since ϵ was arbitrary, we must have $\lim_{N \to \infty} H_N = 0$.

We are now prepared to prove

THEOREM 5. Let $d\infty(x_N) = d(x)$ *Let* $g(x)$ *be a Borel measurable function such that the set* R *of discontinuity points of* $g(x)$ *is closed and* $P(x \subset R) = 0$ *Then* $d\infty[g(x_N)] = d[g(x)]$.

PROOF: Let F_N be the d.f. of x_N, F the d.f. of x, F_{N_g}, F_g the d.f 's of $g(x_N)$, $g(x)$ resp. Then $\lim_{N \to \infty} F_N = F$ in every cont. point of F. Let $h(x)$ be defined as follows:

$$h(x) = 1 \text{ if } g(x) \leq a ,$$

$$h(x) = 0 \text{ if } g(x) > a .$$

The discontinuities of h are contained in the set M of all points where $g(x) = a$ and is continuous or where $g(x)$ is discontinuous. The set R of discontinuity points of $g(x)$ is closed and of measure 0 with respect to F. We can therefore subtract from M a sum R^* of a finite number of open intervals of arbitrarily small measure with respect to F which contains all discontinuity points of $g(x)$. This difference set M' is closed and contains only points where $g(x) = a$ and

$x \notin R$. If a is a continuity point of F_ϱ then the Borel measure of M' with respect to F is 0. Since M' is closed, its Jordan measure is also 0. Hence the Jordan measure of the discontinuity points of $h(x)$ is 0 if a is a continuity point of F_ϱ. Since $g(x)$ is Borel measurable, $\int_{-\infty}^{+\infty} h(x)\, dF_N(x) = F_{N\varrho}(a)$ and $\int_{-\infty}^{+\infty} h(x)\, dF(x) = F_\varrho(a)$ exist for every a. Hence by Lemma 3 $\lim_{N\to\infty} F_{N\varrho}(a) = F_\varrho(a)$ in every continuity point of F_ϱ and this proves our theorem.

3. Corollaries and applications. COROLLARY 2. *If* $\operatorname{plim}_{N\to\infty} (x_N - y_N) = 0$, $d\infty(y_N) = d(y)$ *and if* f *is continuous except in a set* R *for which* $\lim_{N\to\infty} P(y_N \subset R) = 0$ *then* $\operatorname{plim}_{N\to\infty} f(x_N) - f(y_N) = 0$.

PROOF: Let I be a closed interval such that $P(y_N \subset I) \geq 1 - \epsilon/2$. Let I' be a sum of open intervals containing all discontinuity points of $f(x)$ in I and such that $P(y_N \subset I') \leq \epsilon/2$ for sufficiently large N. The set J of points of I which are not points of I' is a closed set. Hence f is uniformly continuous in J and $P(y_N \subset J) \geq 1 - \epsilon$ for sufficiently large N. In Theorem 2 we put $R_N(\epsilon) = J$, $f_N = f$. Then all conditions of Theorem 2 are satisfied and it follows that $\operatorname{plim}_{N\to\infty} [f(x_N) - f(y_N)] = 0$.

If, moreover, the set of discontinuity points of f is closed then by Theorems 3 and 5 $d\infty[f(x_N)] = d\infty[f(y_N)] = d[f(y)]$.

Special cases of Corollary 2 have been proved by J. L. Doob and W. G. Madow (2).

Theorem 5 is very useful in deriving limit distributions.

It follows for instance from Theorem 5 that if $d\infty(x_N) = d(x)$, $d\infty(y_N) = d(y)$, where x, y are independently and normally distributed with mean 0 and equal variances, then $d\infty(x_N/y_N) = d(x/y)$. That is to say the distribution of x_N/y_N converges to a Cauchy distribution.

It also follows from Theorem 5 that under very general conditions the limit distribution of $t = \sqrt{N}(\bar{x} - \mu)/s$ is normal. ($\bar{x}$ = sample mean, μ = population mean, s^2 = sample variance.) For we have under very general conditions $d\infty \sqrt{N}(\bar{x} - \mu) = d(\xi)$, plim $s = \sigma$, where ξ is normally distributed with variance σ^2.

Applying Theorem 5 it can also easily be shown that under very general conditions the limit distribution of T^2 is a chi-square distribution if the means of all variates are 0. Hotelling's T^2 (the generalized Student ratio) for a p-variate distribution is defined as follows:

$$T^2 = N \sum_{i=1}^{p} \sum_{j=1}^{p} A_{ij}\, \xi_i \xi_j \quad \text{where} \quad \| A_{ij} \| = \| s_{ij} \|^{-1}, \qquad \xi_i = \bar{x}^i,$$

where s_{ij} is the sample covariance between x^i and x^j.

We have $d\infty(A_{ij}) = d(\sigma^{ij})$, where $\| \sigma_{ij} \|^{-1} = \| \sigma^{ij} \|$. If $E(x^i) = 0$ for $i = 1, 2, \cdots p$ then $d\infty(\sqrt{N}\, \xi_i) = d(\eta_i)$ where the η_i have a joint normal distribution

with covariance matrix $|| \sigma_{ij} ||$. Hence

$$d \infty (T^2) = d\left[\sum_{i=1}^{p}\sum_{j=1}^{p} \sigma^{ij}\eta_i\eta_j\right] = d\left(\sum_{i=1}^{p} \eta_i'^2\right),$$

where the η_i' are normally and independently distributed with variance 1. Hence the distribution of T^2 converges to a chi-square distribution with p degrees of freedom

If the samples are drawn from a sequence of populations $\{\pi_N\}$ all with the same covariance matrix and such that $\lim_{N\to\infty} \sqrt{N}\mu_{iN} = \mu_i$ where μ_{iN} is the mean value of the ith variate in the Nth population, then one sees in exactly the same way that the limit distribution of T^2 is a non-central square distribution with p degrees of freedom

The limit distribution of T^2 has been derived by W G. Madow (2).

COROLLARY 3. *Let x_N, y_N be r-dimensional vectors $d \infty (y_N) = d(y)$ and $x_N - y_N = O_p[f(N)]$ with $\lim_{N\to\infty} f(N) = 0$. Let $g(x)$ be a function admitting continuous jth derivatives except in a set R with $\lim_{N\to\infty} P(y_N \subset R) = 0$. Let*

$$T_j(x, a) = \sum_{i=1}^{r}\left(\frac{\partial g}{\partial x^i}\right)_{x=a}(x^i - a^i) + \cdots + \left[\sum_{i=1}^{r}(x^i - a^i)\left(\frac{\partial}{\partial x^i}\right)_{x=a}\right]^j g,$$

then

$$g(x_N) - g(y_N) - T_j(x_N, y_N) = o_p\{[f(N)]^j\}.$$

Since the jth derivatives are continuous except in a set of limit measure 0 we can determine a closed set $R(\epsilon)$ on which they are uniformly continuous and so that $P(y_N \subset R(\epsilon)) \geq 1 - \epsilon$ for sufficiently large N. Then for every sequence with $a_N - b_N = O(f(N))$, $b_N \subset R(\epsilon)$ we have

$$g(a_N) - g(b_N) - T_j(a_N, b_N) = o[f(N)^j].$$

Hence Corollary 3 follows from Theorem 1.

Corollary 3 was first proved by W. G. Madow [2] and J. L. Doob [1] for the important case that y_N is a constant.

The following example will illustrate Corollary 3. Let x, y be normally and independently distributed random variables with mean 0 and variance 1; $\{z_N\}$, $\{z_N'\}$ sequences of random variables with $\operatorname{plim}_{N\to\infty} \sqrt{N}\, z_N = \operatorname{plim}_{N\to\infty} \sqrt{N}\, z_N' = 1$. Let $x_N = x + z_N$, $y_N = y + z_N'$. We consider the function $g(x, y) = x^3/3 + y^3/3 + 2x - 2y + 5$. Applying Corollary 1 it is easy to verify that $g(x_N, y_N) - g(x, y) = \Omega_p[1/\sqrt{N}]$, $z_N = O_p(1/\sqrt{N})$, $z_N' = O_p(1/\sqrt{N})$. Hence applying Corollary 3 for $j = 1$ we have

$$g(x_N, y_N) - g(x, y) - (x^2 + 2)z_N - (y^2 - 2)z_N' = o_p(1/\sqrt{N}).$$

Multiplying by $\sqrt{N}$ we have

$$[g(x_N, y_N) - g(x, y)]\sqrt{N} - [(x^2 + 2)z_N + (y^2 - 2)z_N']\sqrt{N} = o_p(1).$$

This is equivalent to

$$\operatorname*{plim}_{N\to\infty}[\sqrt{N}(g(x_N, y_N) - g(x, y))] = x^2 + y^2 .$$

Hence the distribution of $\sqrt{N}(g(x_N, y_N) - g(x, y))$ converges to the chi-square distribution with 2 degrees of freedom.

If $\operatorname*{plim}_{N\to\infty} x_N = a$ and $\{\sigma_N\}$ is a sequence of numbers with $\lim_{N\to\infty} \sigma_N = 0$ such that $d \infty [(x_N^i - a^i)/\sigma_N] = d(\xi_i)$ where the ξ_i are constants or random variables and if g admits continuous first derivatives at $x = a$ at least one of which is different from 0, then putting $\left(\frac{\partial g}{\partial x^i}\right)_{x=a} = g_i$, we have

$$g(x_N) - g(a) = g_1(x_N^1 - a^1) + \cdots + g_r(x_N^r - a^r) + o_p(\sigma_N) .$$

Hence applying Theorems 3 and 5 we have

$$\text{(i)} \qquad d \infty \left[\frac{g(x_N) - g(a)}{\sigma_N}\right] = d(g_1\xi_1 + \cdots + g_r\xi_r).$$

That is to say the distribution of $[g(x_N) - g(a)]/\sigma_N$ converges to the distribution of $\sum_{i=1}^{p} g_i\xi_i$ in all continuity points of the latter. A corresponding result can be obtained from Corollary 3 if all first derivatives are 0 at $x = a$ and at least one second derivative is different from 0 and so forth.

A method of deriving limiting distributions and limit standard deviations based on (i) is known as the δ-method and has been extensively applied in statistical literature.

REFERENCES

[1] J. L. Doob, "The limiting distribution of certain statistics." *Annals of Math. Stat*, Vol. 6 (1935).

[2] W. G. Madow, "Limiting distribution of quadratic and bilinear forms." *Annals of Math Stat*, Vol. 11 (1940).

ON A MEASURE PROBLEM ARISING IN THE THEORY OF NON-PARAMETRIC TESTS

By Henry Scheffé

Princeton University

1. Introduction. While the contents of this paper have broader statistical implications, they were motivated by the following problem: Given two samples, $(Y_1, Y_2, \cdots, Y_m)$ and $(Z_1, Z_2, \cdots, Z_n)$ from univariate populations with cumulative distribution functions (c.d.f's) $F(x)$ and $G(x)$, respectively, and given furthermore that F and G are members of a certain class Ω of c.d.f's, to test the hypothesis that $F = G$. We shall refer to this as "the problem of two samples" [8]. It is an example of what Wolfowitz has called problems of the non-parametric case [8].

For the theory of non-parametric problems the following classification of c.d.f's is appropriate: Let Ω_0 *be the class of all univariate c.d.f's*, that is, the class of all monotone non-decreasing functions $F(x)$ for which $F(-\infty) = 0$, $F(+\infty) = 1$, and $F(x) = F(x.+ 0)$. For every $F \in \Omega_0$ we may conceive of a corresponding random variable X such that $Pr\{X \leq x\} = F(x)$. For some purposes we may desire to rule out the class $\Omega^{(0)}$ of degenerate c.d.f's given by the formula $F(x) = 0$ for $x < x_0$, $F(x) = 1$ for $x \geq x_0$, where x_0 is any real number. Let then Ω_1 *be the class of non-degenerate c d.f's*, $\Omega_1 = \Omega_0 - \Omega^{(0)}$. *Let* Ω_2 *be the class of all continuous* $F(x)$, and let Ω_3 *be the class of all absolutely continuous* $F(x)$, that is, all $F(x)$ for which there exists a probability density function (p.d.f.) $f(x)$ such that

$$F(x) = \int_{-\infty}^{x} f(\xi)\, d\xi. \tag{1}$$

Finally, let Ω_4 *be the class of all* $F(x)$ *which may be expressed in the form* (1) *with* $f(x)$ *continuous.*

Various solutions of non-parametric problems have been given under the restriction that the c.d.f's belong to one of the classes Ω_i. For example, Kolmogoroff [2] has indicated how a confidence belt for an unknown F may be formed with no assumptions on F, that is $F \in \Omega_0$. Wald and Wolfowitz earlier[1] gave a more general solution of the same problem [5], and also of the problem of two samples [6], under the restriction that the c.d f's are members of Ω_2. The latter problem was considered by Dixon [1] for the c.d.f's in Ω_3. Wilks' theory of tolerance intervals [7] assumes $F \in \Omega_4$. The class Ω_1 has been defined above because it is ordinarily the largest class of statistical interest. We note

$$\Omega_0 \supset \Omega_1 \supset \Omega_2 \supset \Omega_3 \supset \Omega_4. \tag{2}$$

[1] See, however, a still earlier paper by Kolmogoroff [11] in which he gave the distribution theory required for his solution.

It is to be understood throughout that the word "*region*" (also the symbol w) *always denotes a Borel set in a k-dimensional* ($k > 1$) *sample space* W (Euclidean). *A "null set" will always mean a Borel set of measure zero*

Returning now to the problem of two samples, let $m + n = k$, $X_i = Y_i$ $(i = 1, 2, \cdots, m)$, $X_i = Z_{i-m}$ $(i = m + 1, \cdots, k)$ Denote by E the point $(X_1, \cdots, X_k)$. Proceeding along the lines of the usual parametric theory, we may seek a region w (the "critical region") such that $Pr\{E \in w\}$ is the same constant α ("significance level"; $\alpha \neq 0$ or 1) for all F in a particular class Ω, if $F = G$. This raises the following question: Define

$$P(w \mid F) = \int_w dF_k(x_1, \cdots, x_k),$$

where

$$F_k(x_1, \cdots, x_k) = \prod_{j=1}^{k} F(x_j).$$

We shall say that a region w *has the property* π_i *if for all* $F \in \Omega_i$, $\alpha = P(w \mid F)$ *is independent of* F *and* $0 < \alpha < 1$. The question then is, for a fixed i, how can we characterize regions w with the property π_i? Partial answers to this question are given in the next section.

In the language of measure theory the question is this: Let μ be any measure on the real line, such that the measure of the whole line is unity, and form the "power" measure μ^k in Euclidean k-space—that is, the product measure obtained by using μ on each axis. For certain large classes C_i (corresponding to the Ω_i defined above, $i = 1, 2, 3, 4$) of measures μ, what can we say about the existence and structure of sets of points in the k-space which have the property that their "power" measure is the same for all measures μ in C_i?

2. Theorems. Our first theorem tells us that if we want regions w with the desired property, we must restrict F to a smaller class than Ω_1.

THEOREM 1: *There is no w with the property* π_1.

To prove the theorem, suppose the contrary. Then there exists a w for which $P(w \mid F) = \alpha$ for all $F \in \Omega_1$ and $\alpha \neq 0$ or 1. Let L be the line $x_1 = x_2 = \cdots = x_k$, and suppose first there is a point E_0 of L in w. Let $E_0 = (a, a, \cdots, a)$, and let $F_h(x)$ be any $F \in \Omega_1$ such that $Pr\{X = a \mid F_h\} = h$ $(0 < h < 1)$. Then

$$\alpha = P(w \mid F_h) \geq P(E_0 \mid F_h) = Pr\{\text{all } X_j = a \mid F_h\}$$

$$= \prod_{j=1}^{k} Pr\{X_j = a \mid F_h\} = h^k.$$

By hypothesis α is independent of h But h may be chosen arbitrarily close to 1. Hence $\alpha = 1$, a contradiction. If no points of w lie on L, the above reasoning applies to $w' = W - w$, since $\alpha' = P(w' \mid F) = 1 - \alpha$ is independent of $F \in \Omega_1$, and w' contains an E_0 on L, therefore $\alpha' = 1$, $\alpha = 0$.

In order to see what kind of structure might yield a w of the desired type, let us,for the moment consider the class Ω_3 of c.d f's Then there exists a p.d.f over W, namely $f(x_1)f(x_2) \cdots f(x_k)$. For any $f(x)$ and any point[2] E, this p d f. has the same value at all points E' whose coordinates are permutations of the coordinates of E. This suggests that suitable regions w can be built up by considering points E for which no two coordinates are equal and putting a fixed fraction of the set $\{E'\}$ in w in such a way that w is a Borel set Our next theorem justifies this process for the wider class Ω_2.

Let us say that *w has the structure S if for every point $E = (x_1, \cdots, x_k)$ with no two coordinates equal, M points $(0 < M < k!)$ of the set $\{E'\}$, obtained by permuting the coordinates of E, are in w and the remaining $k! - M$ are not.*[3]

THEOREM 2: *A sufficient condition that w have the property π_2 is that it have the structure S.*

In proving the theorem it will be convement to separate the $k!$ points of every set $\{E'\}$ by means of regions u_i $(i = 1, \cdots, k!)$, such that each u_i contains one and only one point of $\{E'\}$. Order the $k!$ permutations of the integers $1, 2, \cdots, k$ in any manner so that $(1, 2, \cdots, k)$ is the first. Let $(p_{i1}, \cdots, p_{ik})$ be the ith permutation $(i = 1, 2, \cdots, k!)$ and *define u_i as the region $x_{p_{i1}} < x_{p_{i2}} < \cdots < x_{p_{ik}}$*. The collection $\{u_i\}$ is disjoint and covers all of W except the *set H of points on hyperplanes $x_i = x_j$ $(i \neq j)$* The transformation T_i : $x_{p_{i1}} \to x_1, \cdots, x_{p_{ik}} \to x_k$ maps u_i onto u_1 in such a way that F_k remains invariant.

Suppose now that w satisfies the conditions of the theorem. The removal of $H \cap w$ from w does not[4] affect $P(w \mid F)$ for any $F \in \Omega_2$ Hence

$$P(w \mid F) = \sum_{i=1}^{k!} P(w \cap u_i \mid F) = \sum_{i=1}^{k!} \int_{w \cap u_i} dF_k$$

$$= \sum_{i=1}^{k!} \int_{u_i} c_{w \cap u_i}(E)\, dF_k ,$$

where $c_S(E)$ denotes the characteristic function of a set S, that is, $c_S(E) = 1$ if $E \in S$, 0 otherwise. Next map each of the regions u_i onto u_1 by means of T_i. F_k is invariant, while $c_{w \cap u_i}(E) \to h_i(E)$ such that $\sum_{i=1}^{k!} h_i(E) = M$ for $E \in u_1$. Then

$$P(w \mid F) = \sum_{i=1}^{k!} \int_{u_1} h_i(E)\, dF_k = \int_{u_1} \sum_{i=1}^{k!} h_i(E)\, dF_k = M \int_{u_1} dF_k .$$

---

[2] Previously E denoted a random point $(X_1, \cdots, X_k)$, now it denotes an arbitrary point $(x_1, \cdots, x_k)$ in the sample space W. This will cause no confusion

[3] Regions of structure S may be regarded as the result of applying R A Fisher's randomization process [10] in the most general possible way to the problem of two samples Special cases of regions with structure S have been considered by Feller [9] and Neyman [12], and are implied by all writers [e g., 6] who have attacked the problem of two samples by the method of ranks

[4] This may be seen by writing $P(H \mid F)$ in the form of an integral over W of $c_H(E)\, dF_k$, where $c_H(E)$ is the characteristic function of the set H, and applying the Fubini theorem [4].

But

$$1 = P(W \mid F) = \sum_{i=1}^{k!} \int_{u_i} dF_k ,$$

and by use of T_i we find

$$\int_{u_i} dF_k = \int_{u_1} dF_k \qquad (i = 1, \cdots, k!).$$

Hence

$$\int_{u_1} dF_k = 1/k!,$$

and

$$P(w \mid F) = M/k!$$

for all $F \in \Omega_2$. Thus w has the property π_2.

H is an example of a set in the *class N_2 of regions w for which $P(w \mid F) = 0$ for all $F \in \Omega_2$*. Since if regions w_1 and w_2 differ by a set $w \in N_2$, $P(w_1 \mid F) = P(w_2 \mid F)$ for all $F \in \Omega_2$, we have

COROLLARY 1: *It is sufficient that w have the property π_2 if it differs from a region with structure S by a region in N_2.*

Defining similarly the class N_3 as that class of regions w for which $P(w \mid F) = 0$ for all $F \in \Omega_3$, we see that N_3 is precisely the class of null sets.

COROLLARY 2: *A sufficient condition that w have the property π_3 is that it have the structure S except for a null set.*

The mildest restriction under which the writer has been able to concoct a necessity proof is that the boundary of w be a null set. This class of regions w includes (to the best of his knowledge) all critical regions heretofore used in practice.

THEOREM 3: *For a w whose boundary is a null set, a necessary condition that w have the property π_4 is that it have the structure S except on a null set.*

Suppose then that w has the property π_4, and its boundary B is a null set. Let B_i be the transform of B under T_i. Let the null set H' be the union of H with all B_i, and let $w_1 = w - H'$, $w_2 = (W - w) - H'$. Then w_1 and w_2 are open sets and $P(w_1 \mid F) = P(w \mid F)$ for all $F \in \Omega_4$. Furthermore for any E either all or none of the points of $\{E'\}$ are in $w_1 \cup w_2$. Now consider any $E_0 \in w_1$ and let M_0 be the number of points of $\{E_0'\}$ in w_1, so that $k! - M_0$ of $\{E_0'\}$ are in w_2. Let $E_0 = (\xi_1, \cdots, \xi_k)$, and $2\delta_1 = \min \mid \xi_i - \xi_j \mid$ for $i \neq j$. Since w_1 and w_2 are open, cubes with sides parallel to the coordinate hyperplanes ($x_j =$ constant) and edges of length $2\delta_2$ may be centered on the points E_0' so that each cube is entirely in w_1 or entirely in w_2, by choosing δ_2 sufficiently small. Choose δ so that $\delta > 0$, $\delta < \delta_1$, $\delta < \delta_2$. The set $\{E_0'\}$ is a subset of the set $\{E_0''\}$ of k^k points whose coordinates are in the set $\xi_1, \cdots, \xi_k$ allowing repetitions. For each point $E_0'' = (\xi_{i_1}, \cdots, \xi_{i_k})$ in $\{E_0''\}$ construct a cube $C_{i_1, \cdots, i_k}$ as above

with center at E_0'' and edge 2δ. These cubes are disjoint. Let $f_i(x)$ be a p.d.f. such that the corresponding c.d.f. is in Ω_4 and $f_i(x) = 0$ for $|x - \xi_i| > \delta$ ($i = 1, \cdots, k$). Define the p.d.f.

$$f^{(s)}(x) = s^{-1} \sum_{i=1}^{s} f_i(x) \qquad (s = 1, \cdots, k).$$

Then the corresponding c d.f. $F^{(s)}$ is in Ω_4. We have

$$\alpha = P(w \mid F^{(s)}) = \int_w \prod_{j=1}^{k} f^{(s)}(x_j)\, dW$$

$$= s^{-k} \int_w \sum_{i_1, \cdots, i_k = 1}^{s} f_{i_1}(x_1) \cdots f_{i_k}(x_k)\, dW,$$

where $dW = dx_1 \cdots dx_k$. Bring the last summation sign outside the integral sign, and note that $f_{i_1}(x_1) \cdots f_{i_k}(x_k) = 0$ outside $C_{i_1, \cdots, i_k}$. Then

$$\sum_{i_1, \cdots, i_k = 1}^{s} I_{i_1, \cdots, i_k} = s^k \alpha, \tag{3}$$

where

$$I_{i_1, \cdots, i_k} = \int_{w \cap C_{i_1, \cdots, i_k}} f_{i_1}(x_1) \cdots f_{i_k}(x_k)\, dW. \tag{4}$$

Our argument depends on certain sums of $I_{i_1, \cdots, i_k}$ having the property that the sum is equal to α times the number of terms in the sum. In order to save space we shall say that if Σ is such a sum, then $\Sigma \in R$, R being the class of such sums. Clearly all sums (3) are in R. Let $\{S_{r\nu}\}$ be the subsets of r ($r = 1, \cdots, k$) different integers in the set $1, 2, \cdots, k$ ($\nu = 1, \cdots, {}_kC_r$), and let $\Sigma_{r\nu}$ be the sum of all $I_{i_1, \cdots, i_k}$ for which the index $i_1, \cdots, i_k$ consists only of integers in $S_{r\nu}$ and such that *all* the integers of $S_{r\nu}$ appear in the index. We wish to prove that Σ_{k1}, the sum of I for cubes centered on the points of $\{E_0'\}$, is in R. To accomplish this we make an induction on r: If we assume all $\Sigma_{r\nu} \in R$ for $r < s$, then we can show all $\Sigma_{s\mu} \in R$ ($s = 2, \cdots, k$). No generality is lost in taking $S_{s\mu}$ as the set of integers $1, 2, \cdots, s$. Now consider the left member of (3). Some thought will show[5] that it may be broken down into $\Sigma_{s\mu}$ plus a sum of $\Sigma_{r\nu}$ where $r < s$. But the left member of (3) is in R, and by hypothesis so are all $\Sigma_{r\nu}$ with $r < s$. It follows that $\Sigma_{s\mu}$ is also in R. To see that $\Sigma_{1\nu} \in R$ ($\nu = 1, \cdots, k$), let

[5] To illustrate the reasoning, suppose $s = 4$. If $S_{s\tau}$ is the set of (different) integers $a, b, \cdots, h$, denote $\Sigma_{s\tau}$ by $<a, b, \cdots, h>$, that is, $<a, b, \cdots, h>$ is the sum of all I whose indices contain $a, b, \cdots, h$ and no other integers. Then the right member of (3) contains terms from $<1, 2, 3, 4>$, $<1, 2, 3>$, $<1, 2, 4>$, $<1, 3, 4>$, $<2, 3, 4>$, $<1, 2>$, $<1, 3>$, $<1, 4>$, $<2, 3>$, $<2, 4>$, $<3, 4>$; $<1>$, $<2>$, $<3>$, $<4>$. Every term of the right member of (3) is in one of these sums $< >$. No term can appear in 2 sums $< >$. Every term of each sum $< >$ appears in the right member of (3). Thus the right member is the sum of all sums $< >$ listed above, and by hypothesis, all but the first sum $< >$ are in R.

$S_{1\nu}$ be ν and note that $\Sigma_{1\nu}$ consists only of $I_{\nu,\nu,\ldots,\nu}$. Putting $s = 1$ in (3) we have $I_{1,1,\ldots,1} = \alpha$, and likewise $\Sigma_{1\nu} = I_{\nu,\nu,\ldots,\nu} = \alpha$. Thus $\Sigma_{1\nu} \epsilon R$.

We have at this stage that $\Sigma_{k1} = k!\alpha$. But as we already noted, of the cubes C associated with the integrals I in the sum Σ_{k1}, M_0 are entirely inside w_1 and $k! - M_0$ entirely outside w_1. For the set of M_0 terms in Σ_{k1} corresponding to the cubes C in w_1 the region of integration $w \cap C$ in (4) is actually C, and for the remaining set of terms in Σ_{k1} the region of integration is the empty set. Furthermore if $w \cap C = C$ in (4), the corresponding I is unity. Hence $\Sigma_{k1} = M_0 = k!\alpha$, $\alpha = M_0/k!$. If we now repeated the process with any other point $E_1 \epsilon w_1$ instead of E_0, and let M_1 be the number of points of $\{E_1'\}$ in w_1, we would get $\alpha = M_1/k!$. Therefore $M_1 = M_0$. From $0 < \alpha < 1$, we conclude $0 < M_0 < k!$. Thus w_1 has the structure S.

The exceptional null set allowed for in the statement of Theorem 3 entered the proof when we removed $w \cap H'$ from w. Had we assumed that the boundary $B \epsilon N_2$, then the exceptional set would be in N_2. As a corollary to the reasoning used in the proof we thus get

COROLLARY 3: *If the boundary of w is in N_2, a necessary condition that w have the property π_4 is that w have the structure S except on a subset in N_2.*

Finally, because of (2), any sufficient (necessary) condition for w to have the property π_i is sufficient (necessary) for w to have the property π_j if $j > i$ ($j < i$). Hence we may replace π_2 in Theorem 2 and Corollary 1 by π_3 or π_4, π_3 in Corollary 2 by π_4, π_4 in Theorem 3 and Corollary 3 by π_3 or π_2. This yields

COROLLARY 4: *If the boundary of w is a null set, a necessary and sufficient condition that w have the property π_3 (or π_4) is that it have the structure S except on a null set.*

COROLLARY 5: *If the boundary of w is a region in N_2, a necessary and sufficient condition that w have the property π_2 (or π_3 or π_4) is that it have the structure S except on a subset in N_2.*

3. Remarks. Wald and Wolfowitz [6, 8] in their work on the problem of two samples for the case $F \epsilon \Omega_2$ have imposed the following restriction on any statistic used to test the null hypothesis: The statistic must be a function of V only, where the sequence V of k elements is formed as follows: Rank the X_j of the sample in ascending order of magnitude (ignoring cases where two X_j are equal), and if the i-th element in this rank order is a Y put the i-th element of V equal to zero, else unity. This means that the resulting critical region always consists of the union of s of the regions u_i defined in section 2, where s is a multiple of $m!n!$. The results of our section 2 show that this restriction is not necessary, if all we require is that $Pr\{E \epsilon w\}$, where w is the critical region and E the sample point, be the same constant α whenever the null hypothesis is true. In fact a valid (but probably not very efficient) solution of the problem of two samples has been proposed by Pitman [3] in which the statistic is not a function of V only.

Putting further requirements on the critical region will lead to a more restricted class than the class of regions having essentially the structure S. For instance,

from section 2 it follows that the significance level α can be any of the values $i/k!$ $(i = 1, \cdots, k! - 1)$ But if we lay down a symmetry condition to the effect that if $(y_1, \cdot\ , y_m, z_1, \cdot\ , z_n)$ is in w, all points obtainable by permuting the y's among themselves and the z's among themselves be in w, then α must be a multiple of $m!n!/k!$. Again, if we impose the condition that any statistic $T(X_1, \cdot\ , X_k)$ used to test the null hypothesis remain invariant when all the X_i are subjected to the same topological transformation of the real line onto itself, then Wald and Wolfowitz [6] have shown that T must be a function of V only, so that w has the special structure described above. It would seem desirable when the subject of statistical inference in the non-parametric case may be entering a stage of rapid development, to be clear about the assumptions necessary to restrict the critical region to a particular class.

In concluding these remarks, we quote with the kind permission of Dr Wolfowitz, from some correspondence with the writer. Important work has been done on non-parametric tests under the restriction that the statistic used be invariant under topological transformation The following statement as to *why* this restriction might be imposed will therefore interest the reader· " · there are arguments pro and con · *Pro.* If the statistic be not invariant, this could happen: Two scientists working on the same problem and having the *same* observations to interpret might come to opposite conclusions if one used one scale of measurement and the other used a monotone function of that scale. *Con.* The criterion of topologic invariance of the statistic is a restriction on our freedom. Furthermore it cannot be imposed except in the univariate case ([8], p. 270)."

REFERENCES

[1] W J. Dixon, *Annals of Math Stat*, Vol 11 (1940), pp 199-204

[2] A. Kolmogoroff, *Annals of Math Stat.*, Vol 12 (1941), pp 461–463

[3] E J G Pitman, *J Roy Stat Soc Suppl*, Vol 4 (1937), pp 117–130

[4] S Saks, *Theory of the Integral*, Warsaw, 1937

[5] A Wald and J Wolfowitz, *Annals of Math. Stat.*, Vol 10 (1939), pp. 105–118.

[6] A Wald and J Wolfowitz, *Annals of Math Stat.*, Vol 11 (1940), pp 147–162

[7] S. S. Wilks, *Annals of Math Stat*, Vol 12 (1941), pp. 91–96

[8] J Wolfowitz, *Annals of Math. Stat*, Vol 13 (1942), pp. 247–279.

[9] W. Feller, *Stat Res Mem*, Vol 2 (1938), pp. 107–125

[10] R A. Fisher, *Statistical Methods for Research Workers*, section 24, example 19, *The Design of Experiments*, section 21, *J Roy. Anthrop. Soc*, Vol. 66 (1936), pp 57–63

[11] A Kolmogoroff, *Gior Ist. Ital Attuari*, Vol. 4 (1933), pp 83–91.

[12] J. Neyman, *J Roy Stat Soc*, Vol 105 (1942), pp. 311–312

FURTHER RESULTS ON PROBABILITIES OF A FINITE NUMBER OF EVENTS

By Kai Lai Chung

Tsing Hua University, Kunming, China

In a recent paper[1] the author has generalized some inequalities of Fréchet to the following:

Let $n \geqq a \geqq m \geqq 1$, and let

$$\binom{n-m}{a-m}^{-1} P_a^{(m)}((\nu)) = A_a^{(m)}.$$

$$\Delta F(a) = F(a) - F(a+1), \qquad \Delta^h F(a) = \Delta(\Delta^{h-1} F(a));$$

then

$$\Delta A_a^{(m)} \geqq 0, \qquad \Delta^2 A_a^{(m)} \geqq 0.$$

Using a generalized Poincaré's formula, P. L. Hsu has improved these inequalities to the recurrence formula stated below.

Hsu's formula is

$$\Delta A_a^{(m)} = \frac{m}{n-m} A_{a+1}^{(m+1)}. \tag{1}$$

Proof: We have

$$p_m((\alpha)) = \sum_{b=m}^{a} (-1)^{b-m} \binom{b-1}{m-1} S_b((\alpha)).$$

For a fixed "a" summing over all $(\alpha) \in (\nu)$,

$$\sum_{(a)\in(\nu)} p_m((\alpha)) = \sum_{b=m}^{a} (-1)^{b-m} \binom{b-1}{m-1}\binom{n-b}{a-b} S_b((\nu))$$

$$A_a^{(m)} = \binom{n-1}{m-1} \sum_{b=m}^{a} (-1)^{b-m} \binom{a-m}{b-m}\binom{n-1}{b-1}^{-1} S_b((\nu))$$

$$\Delta A_a^{(m)} = \binom{n-1}{m-1}\left\{\sum_{b=m}^{a} (-1)^{b-m}\left[\binom{a-m}{b-m} - \binom{a+1-m}{b-m}\right]\binom{n-1}{b-1}^{-1} S_b((\nu)) - (-1)^{a+1-m} \cdot \binom{n-1}{a}^{-1} S_{a+1}((\nu))\right\}$$

$$= \binom{n-1}{m-1} \sum_{b=m+1}^{a+1} (-1)^{b-m-1} \binom{a-m}{b-m-1}\binom{n-1}{b-1}^{-1} S_b((\nu))$$

$$= \frac{m}{n-m} A_{a+1}^{(m+1)}, \qquad \text{Q.E.D.}$$

---

[1] "On the probability of the occurrence of at least m events among n arbitrary events," *Annals of Math Stat.*, Vol 12 (1941), pp 328–338. We use throughout the same notation used in this paper, and that referred to in footnote 3

Applying the formula repeatedly, we obtain for $0 \leqq h \leqq n - a$,

$$\Delta^h A_a^{(m)} = \binom{a+m-1}{h}\binom{n-m}{h}^{-1} A_{a+h}^{(m+h)}.$$

Since every $A \geqq 0$, we have, for $0 \leqq h \leqq n - a$,

$$\Delta^h A_a^{(m)} \geqq 0,$$

which includes my former results.

Further, we may write (1) as

$$(n-a)P_a^{(m)} = (a+1-m)P_{a+1}^{(m)} + mP_{a+1}^{(m+1)} \tag{2}$$

or

$$(a+1)P_{a+1}^{(m)} - (n-a)P_a^{(m)} = m\,(P_{a+1}^{(m)} - P_{a+1}^{(m+1)}) = mP_{a+1}^{[m]}$$

It follows that

$$(a+1)P_{a+1}^{(m)} - (n-a)P_a^{(m)} \geqq 0 \tag{3}$$

From (2) it also follows that

$$(n-a)P_a^{(m)} - (a+1-m)P_{a+1}^{(m)} \geqq 0, \tag{4}$$

which is the same as $\Delta A_a^{(m)} \geqq 0$. Combining (3) and (4) we obtain

$$\frac{n-a}{a+1}P_a^{(m)} \leqq P_{a+1}^{(m)} \leqq \frac{n-a}{a+1-m}P_a^{(m)}.$$

If we take the special case $m = 1$ and instead of the original events $E_1, \cdots, E_n$ consider their negations, we easily obtain

$$\frac{n-a}{a+1}\left\{\binom{n}{a} - S_a((\nu))\right\} \leqq \binom{n}{a} - S_{a+1}((\nu)) \leqq \frac{n-a}{a}\left\{\binom{n}{a} - S_a((\nu))\right\}.$$

This is equivalent to a result given by Fréchet[2].

There is an analogue of Hsu's formula for $P_{[m]}$, as follows:
Let $n \geqq a \geqq m \geqq 1$, and let

$$\binom{n-m}{a-m}^{-1} P_a^{[m]} = B_a^{[m]},$$

then

$$\Delta B_a^{[m]} = \frac{m+1}{n-m} B_{a+1}^{[m+1]}.$$

It follows that for $0 \leqq h \leqq n - a$,

$$\Delta^h B_a^{[m]} = \binom{m+h}{m}\binom{n-m}{h}^{-1} B_{a+h}^{[m+h]};$$

$$\Delta^h B_a^{[m]} \geqq 0.$$

[2] "Evénements compatibles et probabilités fictives," *C. R. Acad. Sc.*, Vol. 208 (1939)

The other results on p_m in the paper[1] also have analogues for $p_{[m]}$. For the result on conditions of existence see the author's recent paper[3]. Here we shall state the following extension of Boole's inequality.

For $2l + 1 \leqq n - a$ and $2l \leqq n - a$ respectively, we have

$$\sum_{i=0}^{2l+1} (-1)^i \binom{m+i}{m} S_{m+i}((\nu)) \leqq p_{[m]}((\nu)) \leqq \sum_{i=0}^{2l} (-1)^i \binom{m+i}{m} S_{m+i}((\nu))$$

PROOF. We have

$$S_{m+i}((\nu)) = \sum_{h=0}^{n-m} \binom{m-h}{m+i} p_{[m+h]}((\nu))$$

Hence,

$$\begin{aligned}\sum_{i=0}^{g} (-1)^i \binom{m+i}{m} S_{m+i}((\nu)) &= \sum_{h=0}^{n-m} \left\{ \sum_{i=0}^{g} (-1)^i \binom{m+i}{m} \binom{m+h}{m+i} \right\} p_{[m+h]}((\nu)) \\ &= p_{[m]}((\nu)) + \sum_{h=1}^{n-m} \binom{m+h}{m} \sum_{i=0}^{g} (-1)^i \binom{h}{i} p_{[m+h]}((\nu)) \\ &= p_{[m]}((\nu)) + \sum_{h=1}^{n-m} \binom{m+h}{m} (-1)^g \binom{h-1}{g} p_{[m+h]}((\nu)).\end{aligned}$$

The inequalities follow immediately.

Finally, we record two formulas which express $p_a((\nu))$ in terms of $P_b^{(m)}((\nu))$ and in terms of $P_b^{[m]}((\nu))$ for a fixed m and ranging b's. Formulas which express $P_{[a]}((\nu))$ in both ways have been given[2]

We have,

$$\binom{c-1}{m-1} p((\gamma)) = \sum_{b=m}^{c} (-1)^{b-m} \sum_{(\beta) \in (\gamma)} p_m((\beta))$$

Hence

$$\begin{aligned}\binom{c-1}{m-1} S_c((\nu)) &= \sum_{(\gamma) \in (\nu)} \sum_{b=m}^{c} (-1)^{b-m} \sum_{(\beta) \in (\gamma)} p_m((\beta)) \\ &= \sum_{b=m}^{c} (-1)^{b-m} \binom{n-b}{c-b} \sum_{(\beta) \in (\nu)} p_m((\beta)) \\ S_c((\nu)) &= \binom{c-1}{m-1}^{-1} \sum_{b=m}^{c} (-1)^{b-m} \binom{n-b}{c-b} P_b^{(m)}\end{aligned}$$

By a generalized Poincaré's formula, we get

$$\begin{aligned}p_a((\nu)) &= \sum_{b=m}^{n} (-1)^{b-m} \sum_{c=\max(a,b)}^{n} (-1)^{c-a} \binom{c-1}{a-1} \binom{n-b}{c-b} \binom{c-1}{m-1}^{-1} P_b^{(m)} \\ &= \sum_{b=m+n-a}^{n} (-1)^{n-a+b-m} \binom{b-m}{n-a} \binom{a-1}{m-1}^{-1} P_b^{(m)}.\end{aligned}$$

[3] "On fundamental systems of probabilities of a finite number of events," *Annals of Math. Stat.*, Vol 14 (1943), pp. 123–134

Similarly we have

$$S_a((\nu)) = \binom{c}{m}^{-1} \sum_{b=m}^{c} (-1)^{b-m} \binom{n-b}{c-b} P_b^{[m]}$$

$$p_a((\nu)) = \sum_{b=m}^{n} (-1)^{b-m} \left\{ \sum_{c=\max(a,b)}^{n} (-1)^{c-a} \binom{c-1}{a-1} \binom{n-b}{c-b} \binom{c}{m}^{-1} \right\} P_b^{[m]}$$

It remains to be seen whether the series in the curl brackets can be summed.

Using a formula in footnote 3, we may obtain the desired formula in another way. We have, in fact,

$$\begin{aligned} p_a((\nu)) &= \sum_{c=a}^{m} p_{[c]}((\nu)) \\ &= \sum_{c=a}^{n} \sum_{b=m+n-c}^{n} (-1)^{n-c+b-m} \binom{b-m}{n-c} \binom{c}{m}^{-1} P_b^{[m]}((\nu)) \\ &= \sum_{b=m}^{m+n-a} (-1)^{b-m} \left\{ \sum_{c=m+n-b}^{n} (-1)^{n-c} \binom{b-m}{n-c} \binom{c}{m}^{-1} \right\} P_b^{[m]}((\nu)) \\ &\qquad + \sum_{b=m+n-a+1}^{n} (-1)^{b-m} \left\{ \sum_{c=a}^{n} (-1)^{n-c} \binom{b-m}{n-c} \binom{c}{m}^{-1} \right\} P_b^{[m]}((\nu)). \end{aligned}$$

The "complete" series

$$\begin{aligned} \sum_{c=m+n-b}^{n} (-1)^{n-c} \binom{b-m}{n-c} \binom{c}{m}^{-1} &= \sum_{d=0}^{b-m} (-1)^{d} \binom{b-m}{d} \binom{n-d}{m}^{-1} \\ &= (-1)^{b-m} \frac{m}{n} \binom{n-1}{b-1}^{-1}. \end{aligned}$$

The "incomplete" series we denote by

$$K(n, a, b, m) = \sum_{c=a}^{n} (-1)^{n-c} \binom{b-m}{n-c} \binom{c}{m}^{-1} = \sum_{d=0}^{n-a} (-1)^{d} \binom{b-m}{d} \binom{n-d}{m}^{-1}.$$

Then we may write

$$p_a((\nu)) = \sum_{b=m}^{m+n-a} \frac{m}{n} \binom{n-1}{b-1}^{-1} P_b^{[m]} + \sum_{b=m+n-a+1}^{n} (-1)^{b-m} K(n, a, b, m) P_b^{[m]}.$$

ON THE PROBLEM OF TESTING HYPOTHESES

By R. v. Mises

Harvard University

1. Introduction. The following is known as the problem of testing a simple statistical hypothesis The probability distribution of a variate X depends on a parameter ϑ. In the course of experiments each time a value x of X is observed, one pronounces one of the two assertions: "ϑ equals ϑ_0" or "ϑ is different from ϑ_0." The first assertion is made when the observed value x falls in a "region of acceptance" A, the second, if x falls in the complementary region $\bar{A}$. What is the chance of these assertions being correct and how can A be chosen to make this chance as high as possible?

The distribution for the variate X is considered as given. Let $P(x \mid \vartheta)$ be the probability of the value of X being $\leqq x$ It is obvious that to know $P(x \mid \vartheta)$ is not sufficient for computing the success or error chances of the above assertions. There is another distribution function $P_0(\vartheta)$ involved which we may call the initial or the a priori or the over-all distribution of the parameter ϑ. The meaning of $P_0(\vartheta)$ is as follows. In the infinite sequence of trials there will be among the first N experiences N_1 cases where the assertion that the parameter value is $\leqq \vartheta$ proves correct Then $P_0(\vartheta)$ is the limit of the ratio N_1/N when N tends to infinity. If N_0 is the number of cases in which the actually pronounced assertions $\vartheta = \vartheta_0$ or $\vartheta \neq \vartheta_0$ respectively, prove correct, the limit of N_0/N is the success chance and of $1 - N_0/N$ the error chance of the test under consideration. It would not make any sense to assume that an error chance exists but the over-all chance $P_0(\vartheta)$ does not.[1]

The success and error chances for the assertions $\vartheta = \vartheta_0$ and $\vartheta \neq \vartheta_0$ depend on both functions $P(x \mid \vartheta)$ and $P_0(\vartheta)$ But in most practical cases nothing or very little is known about the parameter distribution. Usually, only the limits within which ϑ varies are known, or a set of distinct values is given which ϑ can assume. Therefore, the problem of testing a hypothesis must be modified in the following way. We ask: *What can be said about the error and success chances of the two alternative assertions and about the choice of the region of acceptance, if $P_0(\vartheta)$ is entirely or partly unknown?* This form of the question corresponds more or less to the conception generally adopted today.

In section 4 of this paper a complete answer to the question is presented for the case of a parameter distribution that is entirely unknown except for the range of possible ϑ-values. This solution, with the restriction to a parameter assuming distinct values only, was already given by Robert W. B Jackson in a paper devoted mainly to some genetical problems [1] The particular circumstances prevailing under the restriction to distinct parameter values will be discussed

[1] The expression "chance" rather than "probability" is used here since no randomness is required Cf the author's paper [2] p 157

in section 8. In section 6 the result is extended to composite hypotheses and in section 7 to problems in several dimensions. An important case of restrictions imposed to $P_0(\vartheta)$ is discussed in section 9.

In the preceding lines the subject of testing a statistical hypothesis was presented in its simplest form, with one scalar variate and one parameter, in order to discard all non-essential complications which would serve only to veil the principal point For the same reason it is to be understood, in the following text, that region (in one dimension) will mean an interval or a finite number of intervals, and distribution will mean a set of concentrated values at distinct points with a continuous density in between or a continuous density throughout If, for the sake of brevity, a Stieltjes integral is used, nothing else is meant than the combination of a sum and an ordinary integral of a continuous function. With respect to the parameter ϑ the distributions $P(x \mid \vartheta)$ are considered as either defined for distinct ϑ-values only or as continuous functions, etc.

2. Error chance. Success rate. J Neyman who must be credited with successfully promoting many problems of mathematical statistics introduced the distinction between errors of first and second type and made this the basis of his approach in dealing with the theory of tests. An error of first kind is committed if the assertion $\vartheta \neq \vartheta_0$ is made when ϑ equals ϑ_0 ; an error of second kind occurs when the assertion $\vartheta = \vartheta_0$ proves incorrect.[2] The chances P_I and P_{II} of these two events can easily be computed, if the distributions $P(x \mid \vartheta)$ and $P_0(\vartheta)$ are considered as known. From $P(x \mid \vartheta)$ we derive the probability $P(A \mid \vartheta)$ for x falling in the region A In particular $P(A \mid \vartheta_0)$ will be designated by $1 - \alpha$. Thus α is the probability of x falling in $\bar{A}$ when $\vartheta = \vartheta_0$. The function $P_0(\vartheta)$ can have, at the point $\vartheta = \vartheta_0$, a jump of magnitude π_0. The set of all ϑ-values except ϑ_0 will be called $\bar{H}$. Then the two error chances are obviously

$$(1) \qquad P_I = \alpha\pi_0 \qquad P_{II} = \int_{(\bar{H})} P(A \mid \vartheta)\, dP_0(\vartheta).$$

By the integral over $\bar{H}$ is meant that the term $P(\bar{A} \mid \vartheta_0)\pi_0$ in the summation has to be omitted The formulae (1) show anew that it would be senseless to speak of error chances without assuming that an over-all distribution $P_0(\vartheta)$ exists.

In all papers that follow Neyman's line of thought first and second type error chances are discussed. But the formulae (1) are seldom written down [3] It is incorrect to say that α is the chance of a first type error and it is likewise incorrect to say that the chance of a second type error depends on ϑ, it depends on the distribution of ϑ.

The total error chance is

$$(2) \qquad P_E = P_I + P_{II} = \alpha\pi_0 + \int_{(\bar{H})} P(A \mid \vartheta)\, dP_0(\vartheta)$$

[2] See e.g. ref [4], [5] or various other publications by the same author

[3] They are included e g in equation (1) of A. Wald's paper [5]

and $1 - P_E$ is the success chance. If the distribution $P(x \mid \vartheta)$, the region of acceptance A, and the test value ϑ_0 are given, P_E depends on $P_0(\vartheta)$ only. If we make $P_0(\vartheta)$ coincide successively with all functions not excluded by some preliminary knowledge about the over-all distribution, there must exist a definite least upper bound (l.u b.) of P_E since P_E has the upper bound 1. The value

$$S = 1 - \text{l.u.b. } P_E$$

is the greatest lower bound of the success chance. In other words, for any positive ϵ there exists a $P_0(\vartheta)$ for which the success chance is $S + \epsilon$ and S is the greatest number for which this holds true We therefore call S the *sure success rate* or, briefly, the *success rate* for the test under consideration. If the success rate S' for a region of acceptance A' is greater than S, the test using A' will be briefly called preferable to that using A.

Neyman's approach consists in comparing two regions A and A' with the same α. The difference of the respective error chances P_E and P'_E is according to (2):

$$P_E - P'_E = \int_{(H)} [P(A \mid \vartheta) - P(A' \mid \vartheta)]\, dP_0(\vartheta) \tag{3}$$

This difference is non-negative, whatever is taken for $P_0(\vartheta)$, if for all values of ϑ

$$P(A \mid \vartheta) \geqq P(A' \mid \vartheta). \tag{4}$$

In this case $P_E \geqq P'_E$ and l.u.b. $P_E \geqq$ l u b. P'_E and therefore $S \leqq S'$. If a region A' can be found for which (4) holds for whatever A, Neyman calls the test using A' a *most powerful test* In fact, this test has at least as large a success rate as any other test using a region of acceptance with the same α Neyman does not use the concept of success rate as introduced here, but implicitly the success chance is the criterion underlying his analysis of tests.[4]

The theory of most powerful tests would supply a complete solution of our problem, if (1) a most powerful test existed in all cases, i.e. for all distributions $P(x \mid \vartheta)$ and all ϑ_0 ; and if (2) a sufficient indication how to chose α were given. Unfortunately it turns out that in almost no practical case a region A' of this kind can be found. The various substitutes for a most powerful test as proposed by Neyman and others (unbiased test, test of type A, etc.) need not be discussed here, since it is obvious that nothing can be said about the difference $S - S'$, if (4) is not fullfilled for all A and ϑ. As to the choice of α, the expression

[4] This can be seen e.g from the justification of most powerful tests as given by A. Wald [7] p 15-16. Moreover, the recommendation of a test with highest success rate as the "best" (which is not the purpose of the present paper) could be justified from the standpoint of the general theory developed by Wald [6] Wald introduces an arbitrary weight function for defining a "best" test If the error weight is taken as one in the case of a false answer and as zero for each correct answer, Wald's "best" test coincides with the test of highest success rate. The present paper includes only statements that refer to the actual numbers of correct and false answers, independently of any arbitrary assumption about an error weight.

"level of significance" used by Neyman, leaves it open whether a high or a low value of α is preferable.

3. Preliminary example. Before attacking the general problem the discussion of a very simple example may provide some information. Let the distribution of the variate X be given by the density

$$p(x \mid \vartheta) = 1 + \vartheta^2(x^2 - \tfrac{1}{3}), \qquad 0 \leqq x \leqq 1. \tag{5}$$

It is immediately seen that the integral of p over the interval 0 to 1 equals 1 for each ϑ and that $p \geqq 0$, if ϑ lies in the limits $-\sqrt{3}$, $\sqrt{3}$. Let this be the only information we possess about the over-all distribution $P_0(\vartheta)$ The value to be tested may be $\vartheta_0 = 0$. The density for this parameter value reduces to $p(x \mid 0) = 1$ and thus the probability of x falling within the interval x_1, x_2 equals $x_2 - x_1$, if $\vartheta = \vartheta_0$. According to the notation introduced above we may consider as intervals of acceptance A all intervals with the limits x_1, $x_1 + 1 - \alpha$, where $0 \leqq x_1 \leqq \alpha$.

The function $P(A \mid \vartheta)$ is now given by

$$\begin{aligned} P(A \mid \vartheta) &= \int_{x_1}^{x_1+1-\alpha} p(x \mid \vartheta)\, dx \\ &= 1 - \alpha + (1 - \alpha)\vartheta^2\left[x_1^2 + x_1(1 - \alpha) - \frac{\alpha(2 - \alpha)}{3}\right]. \end{aligned} \tag{6}$$

In particular, for the interval A' between 0 and $1 - \alpha$:

$$P(A' \mid \vartheta) = 1 - \alpha - (1 - \alpha)\vartheta^2 \frac{\alpha(2 - \alpha)}{3}. \tag{7}$$

The difference of these two expressions is non-negative:

$$P(A \mid \vartheta) - P(A' \mid \vartheta) = (1 - \alpha)\vartheta^2 x_1(x_1 + 1 - \alpha) \tag{8}$$

Thus the interval $0, 1 - \alpha$ is seen to be a most powerful one. The error chance of this test is according to (2):

$$\begin{aligned} P_E' &= \alpha\pi_0 + \int_{(H)}\left[1 - \alpha - \vartheta^2(1 - \alpha)\frac{\alpha(2 - \alpha)}{3}\right]dP_0(\vartheta) \\ &= \alpha\pi_0 + (1 - \alpha)(1 - \pi_0) - (1 - \alpha)\frac{\alpha(2 - \alpha)}{3}\int_{(H)} \vartheta^2\, dP_0(\vartheta). \end{aligned} \tag{9}$$

The last integral is non-negative and can approach zero indefinitely since the total amount $1 - \pi_0$ can be concentrated at a point $\vartheta \neq 0$ with $\vartheta^2 < \epsilon$. Therefore the l.u.b. of P_E' for given α and π_0 is

$$\alpha\pi_0 + (1 - \alpha)(1 - \pi_0)$$

On the other hand, this is a linear function of π_0 which takes its extreme values at the ends of its interval, $\pi_0 = 0$ and $\pi_0 = 1$. Thus the larger of the two values

α and $1 - \alpha$ is the l.u.b. of P'_E, if $P_0(\vartheta)$ is subjected to no further restriction. The success rate of the test under consideration is accordingly the smaller of the two quantities α and $1 - \alpha$.

For $\alpha = 0.99$ or $\alpha = 0.01$ the success rate is 0.01 This means: If we use the most powerful test at a level of significance of either 99% or 1%, we risk in both cases that 99% of all assertions will be false. If $\alpha = \frac{1}{2}$, the success rate reaches its maximum value which is $\frac{1}{2}$ too. On the other hand it can be seen that each interval of length $\frac{1}{2}$ with not too large x_1 would lead to the same success rate. In fact, the error chance P_E for the interval x_1, $x_1 + 1 - \alpha$ is according to (9) and (6)

$$P_E = \alpha\pi_0 + (1 - \alpha)(1 - \pi_0) - (1 - \alpha)\left[\frac{\alpha(2 - \alpha)}{3} - x_1(x_1 + 1 - \alpha)\right]\int_{(\bar{H})} \vartheta^2\, dP_0(\vartheta). \tag{9'}$$

Therefore, the same reasoning as before applies, if the factor in brackets is non-negative. This is the case for $\alpha = \frac{1}{2}$ if the interval begins at a point $x_1 \leqq \frac{1}{4}(\sqrt{5} - 1) = 0.309$ Among these intervals, that with $x_1 = 0$ can be considered as preferable since its success chance for any $P_0(\vartheta)$ is at least as high as that of any other interval.

Now, let us assume that in the definition (5) of $P(x \mid \vartheta)$ the factor ϑ^2 is replaced by some function $g(\vartheta)$ which takes positive and negative values (within $-3/2$ and 3) while ϑ varies from $-\sqrt{3}$ to $\sqrt{3}$. Then equation (6) shows that for any two intervals of acceptance A and A' the difference $P(A \mid \vartheta) - P(A' \mid \vartheta)$ changes its sign at least once with varying ϑ. Thus no most powerful test interval exists. But, applying (9) and calling g_1 the (negative) minimum value of $g(\vartheta)$ we find now

$$\alpha\pi_0 + (1 - \alpha)(1 - \pi_0) - g_1(1 - \alpha)\left[\frac{\alpha(2 - \alpha)}{3} - x_1\,(x_1 + 1 - \alpha)\right](1 - \pi_0)$$

as the l.u.b. of the error chance of A' for given α and π Thus the smaller of the quantities

$$1 - \alpha \quad \text{and} \quad 1 - (1 - \alpha)\left[1 - g_1\frac{\alpha(2 - \alpha)}{3}\right]$$

is the success rate of the test using A'. If g_1 is given we can find, by differentiation the value supplying the highest success rate Using (9') instead of (9) we find in a similar way the success rates for any other interval. It turns out that $S = \frac{1}{2}$ for the interval extending from the above given value $x_1 = 0.309$ to 0.809

There are three things we may learn from this example. (1) It can happen that a most powerful test, at a high or at a low level of significance, has an extremely poor success rate; (2) In the case where a most powerful test with the highest possible success rate exists, there may be other intervals with the same success rate, (3) If no most powerful test exists, there is no need to look

for some substitute definition; the success rate for any kind of test can be found independently of its being most powerful or not.

4. General solution for a simple hypothesis. The distribution $P(x \mid \vartheta)$ of the variate X, the parameter value ϑ_0 to be tested, and the set of all possible values of ϑ are supposed to be given. The set of all possible ϑ-values except ϑ_0 is called $\bar{H}$. Choose a region of acceptance A and compute first, for all ϑ, the magnitude

$$P(A \mid \vartheta) = \int_{(A)} dP(x \mid \vartheta). \tag{10}$$

In particular, the value of this integral for $\vartheta = \vartheta_0$ will be called $1 - \alpha$ and its maximum value or its l.u.b. on $\bar{H}$ will be denoted by β:

$$P(A \mid \vartheta_0) = 1 - \alpha, \qquad \text{l.u.b}_{(\bar{H})}\, P(A \mid \vartheta) = \beta. \tag{11}$$

The chance of committing an error in asserting $\vartheta = \vartheta_0$ when x falls in A or $\vartheta \neq \vartheta_0$ in the case x falls in the complement $\bar{A}$ is according to (2)

$$P_E = \alpha\pi_0 + \int_{(\bar{H})} P(A \mid \vartheta)\, dP_0(\vartheta),$$

where π_0 is the jump of $P_0(\vartheta)$ at the abscissa $\vartheta = \vartheta_0$, or the a priori chance of ϑ_0. The domain of integration over $\bar{H}$ is $(1 - \pi_0)$ and therefore $\beta(1 - \pi_0)$ the l.u.b. of the integral Thus[5]

$$\text{l.u.b. } P_E = \max \{\alpha\pi_0 + \beta(1 - \pi_0)\}.$$

As π_0 can take all values between zero and one, the lowest upper bound of P_E is either α or β. The success rate S, i.e the greatest lower bound of $1 - P_E$, is consequently the smaller of the quantities $1 - \alpha$ and $1 - \beta$.

If the distribution $P(x \mid \vartheta)$ is given and a region of acceptance A for a test value ϑ_0 chosen, the success rate of this test equals the smaller of the two quantities

$$1 - \alpha = P(A \mid \vartheta_0) \quad \textit{and} \quad 1 - \beta = 1 - \text{l.u.b.}_{(\bar{H})}\, P(A \mid \vartheta), \tag{12}$$

if nothing is known about the initial distribution of the parameter except its range. Finding a region of acceptance, A, with the highest success rate, is then a simple maximum-minimum problem.

This solution is not restricted to some rarely occurring type of distributions $P(x \mid \vartheta)$ and it is insofar a complete one as it does not leave undetermined the value of α. Using Neyman's terminology we would have to say: The success rate is the smaller of the two quantities: 1 minus level of significance and minimum power of the test.

It follows from the definitions (12) that, if $P(A \mid \vartheta)$ is continuous in a ϑ-

[5] This formula was given by Jackson [1] p 148 for the "case when the set of alternatives is discontinuous" Jackson calls the test with highest success rate a "most stringent test"

interval including ϑ_0, and ϑ is allowed to take all values of this interval, β cannot be smaller than $1 - \alpha$:

$$\beta \geqq 1 - \alpha \quad \text{or} \quad \alpha + \beta \geqq 1.$$

Thus $1 - \alpha$ and $1 - \beta$ cannot possibly both be greater than $\frac{1}{2}$. The greatest possible success rate is then $\frac{1}{2}$ and it can be reached only if $\alpha = \beta = \frac{1}{2}$. We state: *No test can have a success rate S greater than $\frac{1}{2}$, if ϑ can vary in an interval including ϑ_0 without any restriction and $P(A \mid \vartheta)$ is a continuous function of ϑ in this interval.*

We will see later, in sections 8 and 9, how certain restrictions imposed to $P_0(\vartheta)$ which are effective in some problems improve the success rate of a test.

5. Examples. Let us assume that the variate X is normally distributed according to

$$P(x \mid \vartheta) = \Phi[h(x - \vartheta)], \qquad \Phi(u) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{u} e^{-x^2}\, dx. \tag{13}$$

The parameter value to be tested may be taken as $\vartheta_0 = 0$ without loss of generality, since in all other cases $X - \vartheta_0$ can be considered as the variate. If the interval x_1, x_2 is chosen for the region of acceptance, we have

$$P(A \mid \vartheta) = \phi[h(x_2 - \vartheta)] - \phi[h(x_1 - \vartheta)]. \tag{14}$$

The right hand side becomes a maximum, if

$$\phi'[h(x_2 - \vartheta)] = \phi'[h(x_1 - \vartheta)], \qquad \text{i.e.} \quad \vartheta = \tfrac{1}{2}(x_1 + x_2).$$

Therefore, for $\vartheta_0 = 0$

$$1 - \alpha = \phi(hx_2) - \phi(hx_1), \qquad \beta = \phi(\tfrac{1}{2}h(x_2 - x_1)) - \phi(\tfrac{1}{2}h(x_1 - x_2)).$$

Both quantities have the value $\frac{1}{2}$, if and only if

$$x_1 = -x_2, \qquad \phi(hx_1) = \tfrac{1}{4}, \qquad \phi(hx_2) = \tfrac{3}{4}. \tag{15}$$

These are the *probable limits* of x. The conclusion is that *the probable limits supply the interval with the highest possible success rate $S = \frac{1}{2}$.*

The result is not restricted to the particular form of the function ϕ, it remains valid, if ϕ is replaced by any function whose derivative ϕ' has one maximum and decreases both ways symmetrically. It is well known that this test which has always been used by statisticians and is here proved to have the maximum success rate, is neither most powerful nor even, for a general ϕ, unbiased. We also see that the interval determined by (15) is the only closed interval with maximum success rate.

Our method supplies the analogous solution for the case of an unsymmetric distribution also. Assume the density

$$p(x \mid \vartheta) = f(x - \vartheta), \tag{16}$$

where $f(u)$ is supposed to have only one maximum, say at the point $u = 0$. The value to be tested may again be chosen as $\vartheta_0 = 0$. For the interval x_1, x_2 as region of acceptance we have

$$P(A \mid \vartheta) = \int_{x_1}^{x_2} f(x - \vartheta)\, dx = \int_{x_1-\vartheta}^{x_2-\vartheta} f(u)\, du.$$

The last expression becomes a maximum with respect to ϑ, if

$$f(x_1 - \vartheta) = f(x_2 - \vartheta).$$

The maximum will occur at the point $\vartheta = 0$ and accordingly coincide with $1 - \alpha$, if $f(x_1) = f(x_2)$. Thus we have a region of acceptance with the highest possible success rate $\frac{1}{2}$, if x_1, x_2 are determined by

$$\int_{x_1}^{x_2} f(u)\, du = \tfrac{1}{2}, \qquad f(x_1) = f(x_2). \tag{17}$$

Under the assumptions made for $f(u)$ there exists exactly one pair of values x_1, x_2 obeying these equations. This kind of test too has been much used by statisticians, but an account of its merits has so far not been given

Another example is supplied by the density function

$$p(x \mid \vartheta) = \vartheta^2 x e^{-\vartheta x}, \qquad x \geqq 0, \quad \vartheta > 0 \tag{18}$$

We derive for an interval x_1, x_2

$$P(A \mid \vartheta) = \int_{x_1}^{x_2} p(x \mid \vartheta)\, dx = (\vartheta x_1 + 1)e^{-\vartheta x_1} - (\vartheta x_2 + 1)e^{-\vartheta x_2}.$$

If ϑ_0 is the value to be tested, we have

$$1 - \alpha = (\vartheta_0 x_1 + 1)e^{-\vartheta_0 x_1} - (\vartheta_0 x_2 + 1)e^{-\vartheta_0 x_2}. \tag{19}$$

One may ask for an interval x_1, x_2 with the success rate $S = \frac{1}{2}$ Then equation (19) must be fulfilled with $\alpha = \frac{1}{2}$ and, moreover, $P(A \mid \vartheta)$ must take its maximum value at $\vartheta = \vartheta_0$. This provides the second condition

$$\frac{\partial P(A \mid \vartheta)}{\partial \vartheta} = 0 \quad \text{at} \quad \vartheta = \vartheta_0, \quad \text{i.e.} \quad x_2^2 e^{-\vartheta_0 x_2} = x_1^2 e^{-\vartheta_0 x_1}. \tag{19'}$$

There exists, for each $\vartheta_0 > 0$, one and only one pair of values x_1, x_2 obeying the two equations (19) and (19').

In all these examples it turned out that at least one interval with the success rate $S = \frac{1}{2}$ (the highest value for a distribution continuous with respect to ϑ) exists. It seems that this is a common property of most usual distribution functions $P(x \mid \vartheta)$. But we can easily give an example where the greatest S, at least for a single interval as region of acceptance, is smaller than $\frac{1}{2}$. Assume

$$P(x \mid \vartheta) = x + \vartheta x(1 - x)(2\vartheta^2 x - 1), \qquad 0 \leqq x \leqq 1, \qquad -1 \leqq \vartheta \leqq 1, \tag{20}$$

and let $\vartheta_0 = 0$ be the value subjected to testing. For any interval beginning at x and extending to $x + 1 - \alpha$ we find

$$P(A \mid \vartheta) = 1 - \alpha + a\vartheta + b\vartheta^3 \quad \text{with} \quad a = (1-\alpha)(2x-\alpha), \tag{21}$$
$$b = 2(1-\alpha)(-3x^2 + 3\alpha x - \alpha^2 + \alpha - x).$$

It is a necessary condition for a test with $S = \frac{1}{2}$—in the case of a differentiable $P(A \mid \vartheta)$—that the derivative of $P(A \mid \vartheta)$ vanishes at $\vartheta = \vartheta_0$. Thus we must have

$$\frac{\partial P(A \mid \vartheta)}{\partial \vartheta} = a + 3b\vartheta^3 = 0 \quad \text{for} \quad \vartheta = 0. \tag{22}$$

This shows that $2x - \alpha$ must be zero or $x = \frac{1}{2}$. On the other hand, for $\alpha = \frac{1}{2}$, $x = \frac{1}{4}$ the formula for $P(A \mid \vartheta)$ becomes

$$P(A \mid \vartheta) = \tfrac{1}{2} + \tfrac{3}{16}\vartheta^3.$$

Thus P has an inflexion point at $\vartheta = 0$ and its maximum, β, must be greater than $\frac{1}{2}$. In the present example, as ϑ goes up to 1, we have $\beta = 11/16$ and the success rate is $S = 5/16$. This does not exclude that intervals with a success rate between $5/16$ and $\frac{1}{2}$ exist. E.g. for $x = 0.45$ and $\alpha = \frac{1}{2}$ one finds the maximum $\beta = 0.60$ and thus $S = 0.40$. The optimum interval can be found by differentiating the formula for $P(A \mid \vartheta)$ with respect to x and α.

Examples with the ϑ restricted to distinct values will be discussed in section 8.

6. Composite hypotheses. We have the problem of testing a composite hypothesis, if instead of one value ϑ_0 a region H of ϑ-values is given and the assertions to be made in the course of experiments are "ϑ belongs to H" or "ϑ does not belong to H." The solution developed in section 4 applies to this case almost without modification.

Again, let $P(A \mid \vartheta)$ be the probability of x falling in the region of acceptance A. By $\bar{A}$ and $\bar{H}$ we denote the regions complementary to A in the sample space and to H in the ϑ-space. Then the error chance is

$$P_E = \int_{(H)} [1 - P(A \mid \vartheta)]\, dP_0(\vartheta) + \int_{(\bar{H})} P(A \mid \vartheta)\, dP_0(\vartheta). \tag{23}$$

This is an obvious generalisation of (2). The equation expresses the fact that each time x falls in $\bar{A}$ and ϑ in H or x in A and ϑ in $\bar{H}$, an error is committed. Let us use the notations

$$\pi_0 = \int_{(\bar{H})} dP_0(\vartheta)$$
$$\alpha = \text{l.u.b. of } P(\bar{A} \mid \vartheta) \text{ for } \vartheta \text{ in } H \tag{24}$$
$$\beta = \text{l.u.b. of } P(A \mid \vartheta) \text{ for } \vartheta \text{ in } \bar{H}$$

Then the first of the two integrals in (22) cannot be greater than $\alpha\pi_0$ and the second not greater than $\beta(1 - \pi_0)$. On the other hand no lower upper bound exists for either of these integrals, if π_0 is given and $P_0(\vartheta)$ subjected to no other restriction.

As π_0 varies between 0 and 1, the expression

$$\alpha\pi_0 + \beta(1 - \pi_0)$$

has its extreme values at the points $\pi_0 = 0$ and $\pi_0 = 1$ and these values are α and β. Accordingly the greater of the quantities α and β is the l.u.b. of P_E and the success rate S equals the smaller of the two quantities $1 - \alpha$ and $1 - \beta$. If $P(A \mid \vartheta)$ is continuous with respect to ϑ, we have again $\beta \geqq 1 - \alpha$, thus α and β cannot be both smaller than $\frac{1}{2}$ and no S can become $> \frac{1}{2}$.

If the hypothesis that ϑ lies in H is tested by means of a region of acceptance A, the success rate of this test equals the smaller of the two quantities $1 - \alpha$ and $1 - \beta$ which are the minimum of $P(A \mid \vartheta)$ for ϑ-values in H and the minimum of $P(\bar{A} \mid \vartheta)$ for ϑ-values outside H. The task of finding the region A with highest success rate is thus reduced to a simple maximum-minimum problem.

As an example let us take the density function

$$p(x \mid \vartheta) = f(x - \vartheta), \tag{25}$$

where $f(u)$ has a maximum at $u = 0$ and drops on both sides symmetrically and monotonically towards zero. The hypothesis to be tested may be given as

$$-b \leqq \vartheta \leqq b.$$

We find, if the interval x_1, x_2 is taken for region of acceptance:

$$P(A \mid \vartheta) = \int_{x_1}^{x_2} f(x - \vartheta)\, dx = \int_{x_1-\vartheta}^{x_2-\vartheta} f(u)\, du. \tag{26}$$

This function of ϑ has its maximum at $\vartheta = \frac{1}{2}(x_1 + x_2)$ and drops symmetrically both sides. If $\frac{1}{2}(x_1 + x_2)$ is supposed to lie in the interval $(0, b)$ we find

$$1 - \alpha = \int_{x_1+b}^{x_2+b} f(u)\, du, \qquad \beta = \int_{x_1-b}^{x_2-b} f(u)\, du$$

Both quantities reach the value $\frac{1}{2}$, if we choose $x_2 = -x_1 = a$ and take for a the uniquely determined solution of

$$\int_{a+b}^{a+b} f(u)\, du = \int_{-a-b}^{a-b} f(u)\, du = \tfrac{1}{2}. \tag{27}$$

For this interval the success rate has its highest possible value $\frac{1}{2}$.

7. Case of n variates and k parameters. The analysis given in section 4 for a simple hypothesis and in 6 for a composite one extends immediately to the case where instead of one variate X and one parameter ϑ a group of n variates $X_1, X_2, \cdots, X_n$ and a group of k parameters $\vartheta_1, \vartheta_2, \cdots, \vartheta_k$ are in question.

The region of acceptance A is now a portion of the n-dimensional sample space, determined by an interval of a function $F(x_1, x_2, \cdots x_n)$. The hypothesis to be tested will consist in assuming that the point $\vartheta_1, \vartheta_2, \cdots \vartheta_k$ falls into a certain region H of the k-dimensional parameter space. The success rate of such a test is again the smaller of the numbers $1 - \alpha$ and $1 - \beta$ where α and β are defined in exactly the same way as in the preceding section. The minimum of $P(A \mid \vartheta)$ when the ϑ-values fall into H is called $1 - \alpha$, and the maximum of the same function for all ϑ-combinations belonging to the complementary region $\bar{H}$ is β.

If the test function $F(x_1, x_2, \cdots x_n)$ is known, the interval with the highest success rate, can be found on the same lines as in the case of one variate. In fact, the quantity F takes the place of x in the former analysis. If the interval thus found has the success rate $\frac{1}{2}$, we know that no other test exists which would have a higher success rate as long as nothing is known about the a priori distribution in the parameter space. If a certain $F(x_1, x_2, \cdots x_n)$ does not lead to an interval with success rate $\frac{1}{2}$, one may try another test function. In the most general case the test function F with the highest success rate would be found by solving the problem of calculus of variation that consists in maximizing $1 - \alpha$ and $1 - \beta$. As a rule such an elaborate analysis will not be necessary.

To ask that a test be a most powerful one is too much and too little. It is too much since such a test does not exist in most cases. It is too little because there can exist another test (on a different level of significance) with a considerably higher success rate. The correct description of a most powerful test is that such a test can be shown, in a simple way, to have no smaller success chance whatever $P_0(\vartheta)$ is than a group of other tests. If a most powerful test exists, it may be considered preferable to all other tests of the same success rate, but there is no reason why it should be considered more favorable than any test with higher success rate. As to unbiased tests, and other substitutes for most powerful tests, nothing at all can be said about their merits as compared with that of other tests.

A simple example for tests with the highest possible success rate in the case of several dimensions is the following. Assume a density function

$$p(x \mid \vartheta) = f(x_1 - \vartheta_1, x_2 - \vartheta_2, \cdots x_n - \vartheta_n) \tag{28}$$

where $f(u_1, u_2, \cdots u_n)$ depends on the absolute values $\mid u_1 \mid, \mid u_2 \mid, \cdots \mid u_n \mid$ only and decreases monotonically with increasing $u_1^2 + u_2^2 + \cdots u_n^2$ in all directions. The parameter point $\vartheta_1 = \vartheta_2 = \cdots \vartheta_n = 0$ is to be tested. Let $F(x_1, x_2 \cdots x_n)$ be a function likewise depending on $\mid x_1 \mid, \mid x_2 \mid, \cdots \mid x_n \mid$ only, vanishing at the origin, and monotonically increasing with $x_1^2 + x_2^2 + \cdots x_n^2$. Then the set of points for which

$$F(x_1, x_2, \cdots x_n) \leqq C \tag{29}$$

is a region of acceptance with success rate $\frac{1}{2}$, if C is chosen in such a way as to have

$$\int_{(F \leqq C)} f(x_1, x_2, \cdots x_n)\, dx_1\, dx_2 \cdots dx_n = \tfrac{1}{2}. \tag{30}$$

This applies e.g. to normal populations. The proof is obvious.

8. Distinct parameter values. Tests with higher success rate than $\frac{1}{2}$ can be found, if the parameter ϑ is restricted to a set of distinct values. Take for instance our first example in section 3 and assume that ϑ can only take the three values 0, ± 1. Then in the second expression (9) for the error chance the integral can not approach the value zero since the region $\bar{H}$ does not include the point $\vartheta = 0$. The minimum value of the integral is $(1 - \pi_0)$ and thus

$$P'_E \leqq \alpha\pi_0 + (1 - \alpha)\left[1 - \frac{\alpha(2 - \alpha)}{3}\right](1 - \pi_0). \tag{31}$$

The success rate is the smaller of the two quantities

$$1 - \alpha \quad \text{and} \quad 1 - (1 - \alpha)\left[1 - \frac{\alpha(2 - \alpha)}{3}\right] = 1 - \beta.$$

The best value of α is found by equating α and β. This gives about $\alpha = \beta = 0.436$ and the success rate $S = 0\,564$, for the region of acceptance $x = 0$ to $x = 0.564$. Other intervals or sets of intervals can be examined in the same way

A more impressive example is the following. We draw $n = 12$ times from an urn which contains three balls, black ones and white ones. The observed value x is the number of white balls drawn. The probability ϑ of getting a white ball in one experiment can have one of the four values 0, 1/3, 2/3, 1, and we want to test the hypothesis $\vartheta = \vartheta_0 = 1/3$. The probability distribution is given by

$$\pi(x \mid \vartheta) = C_n^x \vartheta^x (1 - \vartheta)^{n-x} \tag{32}$$

Let us choose the set of points $x = 1, 2, \cdots 6$ as region of acceptance. Then

$$P(A \mid \vartheta) = \sum_{x=1}^{6} C_n^x\, \vartheta^x (1 - \vartheta)^{n-x}. \tag{33}$$

This sum can be computed for the 4 possible ϑ-values:

$P(A \mid \vartheta) =$	0	0.926	0.178	0
for $\vartheta =$	0	1/3	2/3	1

Thus $1 - \alpha$ has the value 0.926 and β equals 0.178 The success rate is the smaller of the two quantities 0 926 and 0.822, thus $S = 0.822$. If we restrict the region of acceptance to the points $x = 1$ to 5, the values of $1 - \alpha$ and $1 - \beta$ become 0.815 and 0.934, thus the success rate $S = 0.815$. In the first case we have more than 82% chance of making a correct assertion, whatever the a priori probability of ϑ may be!

It is obvious that this result will become more and more strongly marked, if the number of observations increases. This is connected with the subject of the next section.

9. Asymptotically increasing success rate. It seems strange that in the case of a continuously varying parameter and a distribution $P(x \mid \vartheta)$ which is continuous with respect to ϑ no test can have a success rate $> \frac{1}{2}$. One has the feeling that something might happen in the continuous problems similar to what was the case in the example of section 8 On the other hand our proof that $S \leqq \frac{1}{2}$, in sections 4 and 6, is conclusive and it applies to problems in more than 1 dimension also. The answer is that in the kind of problems where a large number of observations is involved a definite restrictive assumption about the over-all distribution $P_0(\vartheta)$ is silently introduced.

The problems we have here in mind are connected with sequences of distributions of the form

$$P_n(x \mid \vartheta) = \phi_n(x - \vartheta), \tag{34}$$

where $\phi_1(u)$, $\phi_2(u)$, $\phi_3(u)$, $\cdots$ are cumulative distribution functions for distributions more and more concentrated around one point, say $u = 0$. In a rigorous form the sequence $\phi_n(u)$ can be described by the following statement: For each $\epsilon, \eta > 0$ exists a number $N(\epsilon, \eta)$ such that

$$\phi_n(\eta) - \phi_n(-\eta) \geqq 1 - \epsilon \qquad \text{for } n > N(\epsilon, \eta). \tag{35}$$

One wants to test the hypothesis

$$-b \leqq \vartheta \leqq b,$$

under the assumption that *the parameter distribution does not depend on n*. In this case, as we shall show, one can find for each $\epsilon > 0$ a region of acceptance A such that the success rate S_n of the test corresponding to this A and to $P_n(x \mid \vartheta)$ is greater than $1 - \epsilon$ for sufficiently large n.

We divide the region $\bar{H}$, i.e. $\mid \vartheta \mid > b$, into two parts $\bar{H}_1$ and $\bar{H}_2$ where $\bar{H}_1$ consists of the points $\mid \vartheta \mid \leqq b + 2\eta$ and satisfies the condition

$$\int_{(\bar{H}_1)} dP_0(\vartheta) \leqq \frac{\epsilon}{3}. \tag{36}$$

Then the region of acceptance will be

$$-a = -b - \eta \leqq x \leqq b + \eta = a,$$

and the probability of x falling in this region:

$$P_n(A \mid \vartheta) = \phi_n(b + \eta - \vartheta) - \phi_n(-b - \eta - \vartheta). \tag{37}$$

As long as ϑ belongs to H the right hand side in (37) is not smaller than $\phi_n(\eta) - \phi_n(-\eta)$ and thus, according to (35) the error chance of first kind

$$P_I^{(n)} = \int_{(H)} [1 - P_n(A \mid \vartheta)]\, dP_0(\vartheta) \leqq 1 - [\Phi_n(\eta) - \Phi_n(-\eta)] \leqq \frac{\epsilon}{2} \tag{38}$$

$$\text{for } n > N\left(\frac{\epsilon}{3}, \eta\right).$$

The error chance of second kind can be written as

$$P_{II}^{(n)} = \int_{(\bar{H}_1)} P_n(B \mid \vartheta)\, dP_0(\vartheta) + \int_{(\bar{H}_2)} P_n(A \mid \vartheta)\, dP_0(\vartheta). \tag{39}$$

The first of these integrals cannot be larger than $\frac{\epsilon}{3}$ according to (36) since $P_n(A \mid \vartheta) \leqq 1$. The second integral cannot exceed the maximum value of $P_n(A \mid \vartheta)$ for ϑ in $\bar{H}_2$. But if $\mid \vartheta \mid > b + 2\eta$ the two arguments of ϕ_n in (37) have always the same sign and are in absolute value greater than η. It then follows from (35), in connection with the fact that $\phi_n(u)$ increases monotonously from 0 to 1, that the difference of the two ϕ_n-values cannot exceed $\frac{\epsilon}{3}$ for $n > N(\epsilon/3, \eta)$. Therefore

$$P_{II}^{(n)} \leqq \frac{\epsilon}{3} + \frac{\epsilon}{3} \text{ and } S_n = 1 - P_I^{(n)} - P_{II}^{(n)} \geqq 1 - \epsilon \text{ for } n > N\left(\frac{\epsilon}{3}, \eta\right). \tag{40}$$

This result has a wide range of application in the cases where a hypothesis is tested on the basis of a large number of independent observations. Consider a sequence of variates X_1, X_2, X_3, $\cdots$ subject to probability distributions $Q_1(x_1)$, $Q_2(x_2)$, $Q_3(x_3)$, $\cdots$ Let $x = F(x_1, x_2, \cdots x_n)$ be a *statistical function*, i.e. a function depending on the *distribution* of its n variables only, and ϑ the expected value of F. Then the general law of large numbers states that the distribution of x has the form (34) with ϕ_n satisfying the inequality (35), if the $Q_n(x)$ fulfill certain conditions concerning mainly their behaviour at infinity[6]. The proof of this theorem which is the real source of most "asymptotical" properties of statistical tests was given for the first time in 1936 The particular case where F is the arithmetical mean of the n variables $x_1, x_2, \cdots x_n$ has been known as Tchebychef's theorem since 1867

Applying this general law of large numbers we can now state the following fact. *In testing a hypothesis about the expected value ϑ of any regular statistical function of n variates we can reach a success rate $1 - \epsilon$, no matter how small ϵ is, if the number n increases indefinitely and the initial distribution of ϑ is supposed to be independent of n* On the other hand, no test with a success rate greater than $\frac{1}{2}$ is available, if an assumption of this type is not used.

[6] For exact conditions see ref [3].

10. Summary. In this paper a solution of the problem of testing hypotheses is presented in the following sense. It is assumed that a probability distribution depending on some parameters is given and that nothing is known about the initial distribution of these parameters. For any simple or composite hypothesis about the parameters and any region of acceptance chosen in the sample space the success rate S is computed, i.e. the minimum chance for getting right answers out of the test. From the formulae given for S a test with highest success rate can easily be found in each case.

This theory shares the point of departure with the actually used theory which leads to the concept of most powerful tests. A most powerful test is described as a test which, by simple reasoning, can be seen to have no smaller success chance than any other test on the same "level of significance" α. In the rare cases where most powerful tests exist for all α-values, one of them, with an α-value singled out by our theory, has the highest success rate and then is preferable to all other tests which might have the same success rate. In all other cases our method supplies a test of highest success rate in no relation to "unbiased" tests or other current substitutes for most powerful tests.

Some of the main results are: No test has a success rate $>\frac{1}{2}$, if nothing is known about the parameters except the limits of their values and if the given distribution is a continuous function of the parameters. The success rate can be higher, if the parameters are restricted to certain distinct values. A success rate no matter how close to 1 can be reached in a sequence of tests based on an increasing number n of observations, if the initial distribution of the parameters is known to be independent of n.

REFERENCES

[1] ROBERT W. B. JACKSON, "Tests of statistical hypotheses in the case when the set of alternatives is discontinuous, illustrated on some genetical problems." *Stat Res. Mem.*, Vol 1 (1936), p 138-161

[2] R. v MISES, "On the correct use of Bayes' formula," *Annals of Math Stat*, Vol 13 (1942), p. 156-165

[3] R. v. MISES, "Die Gesetze der grossen Zahl für statistische Funktionen," *Monatsh Mathem u. Physik*, Vol 43 (1936), p. 105-128

[4] J NEYMAN, "Sur la vérification des hypothèses statistiques composées," *Bull. Soc Math de France*, Vol. 63 (1935), p 246-266.

[5] J. NEYMAN, "Outline of a theory of statistical estimation based on the classical theory of probability," *Phil Trans.*, Ser A, Vol 236 (1937), p. 333-380

[6] A. WALD, "Contributions to the theory of statistical estimation and testing hypotheses," *Annals of Math. Stat.*, Vol 10 (1939), p 299-326

[7] A WALD, "On the principles of statistical inference," 1942, *Notre Dame Lect* No 1

ON THE RELIABILITY OF THE CLASSICAL CHI-SQUARE TEST

By E. J. Gumbel

New School for Social Research

For a given set of observations and for a continuous variate, different classifications lead to different observed distributions and to different values of χ^2. This shortcoming has been vaguely felt by statisticians. We shall explain how these differences arise and show that they are important enough to cast a great deal of doubt on the validity of the application of the usual χ^2 method to a continuous variate. Finally, we propose a procedure which is free from these difficulties.

1. The observed distributions. The χ^2 method gives a numerical measure of the differences between the observed and the theoretical distribution. A theoretical distribution is completely determined once the constants are known. For a discontinuous variate the observed distribution is also well defined; but for a continuous variate the concept "observed distribution" is vague. To classify N observations, $x_1, x_2, \cdots x_m, \cdots x_N$ arranged in increasing order, we introduce two arbitrary actions: the choice of the intervals and the beginning of the first cell As a rule, all cells have the same length, and they are bounded by integral numbers, or even numbers, or round numbers, 0, 5, 10, of the variate. But these classifications and the preference given to round numbers for the starting point have no theoretical foundation.

A certain guide for the systematic choice of the class length and the beginning of the first cell may be found by turning to the theory. Many theoretical distributions of a continuous variate x have only two constants, and permit the introduction of a reduced variate y with the dimension zero, where

$$y = \frac{x - a}{b}. \tag{1}$$

The constant a is a mean, and b is a measure of dispersion. The probabilities $W(x)$ (or $F(y)$) for values equal to or less than x (or y) are

$$W(x) = F(y). \tag{2}$$

For most distributions, for which the above transformation is possible, tables for $F(y)$ exist, in which the argument progresses by a fixed interval Δy. By taking an initial value y_0 and a fixed interval Δy, the differences

$$NF(y_0 + i\Delta y) - NF(y_0 + (i-1)\Delta y) = Np_i \qquad (i = 1, 2, \cdots k) \tag{3}$$

may be interpreted as being the theoretical distribution. The corresponding values of the variate, by (1), are

$$x(i) = a + b(y_0 + i\Delta y); \qquad x(i-1) = a + b(y_0 + (i-1)\Delta y) \tag{4}$$

and the cell length is

$$\Delta(x) = b\Delta y. \tag{5}$$

In (3) k is the number of cells. In general, $x(i)$ and $x(i-1)$ will not exist among the observed values x_m. By arranging the observations in the cells given by the theoretical values (4), we obtain an observed distribution consisting of the contents a_i of the cell i. This procedure prescribes a classification of the observations according to the theory. The intervals selected are multiples of some measure of dispersion. In principle, the choice of Δy and of the starting point y_0 remain arbitrary; in practice, the selection of Δy is limited by the intervals given in the probability tables.

This natural classification may be used for constructing *different* observed distributions from the same set of observations. We determine the constants, then choose a small interval and a starting point which is below the smallest observation x_1. The last cell is such that it contains the largest observation x_N. In this way, we obtain the initial observed distribution, consisting of k cells

If we combine h cells ($h = 2, 3, \cdots \frac{1}{2}k$), we obtain h different observed distributions: We combine $h-1$ void cells with the first cell of the initial distribution, we combine the second cell and the following $h-1$ cells of the initial distribution, and so on. Generally, we combine q void cells ($q = h-1, h-2, \cdots 1, 0$) with the first $h-q$ cells of the initial distribution, then the next h cells of the initial distribution, and so on The last of these h distributions starts with the first h cells of the initial distribution.

If we combine more and more cells, the number of observed distributions, having the same intervals, increases. The larger the intervals the larger is the influence of the starting point, and the more the observed distributions become dissimilar. To see this influence of classification on the shape of the observed distributions, consider the extreme case for a symmetrical theoretical distribution of an unlimited variate. Let the observed distribution consist of two cells Assume besides that the observed median is close to the theoretical one. If the cut between the cells is identical with the theoretical median, the two cells have the contents $\frac{1}{2}N + \epsilon$ and $\frac{1}{2}N - \epsilon$, where ϵ is small. If the cut is shifted sufficiently far to the left or right of the median, the cell contents will be 0, N and N, 0. These two distributions are completely different.

To each observed distribution corresponds a theoretical one obtained from (3) by the same combination of cells as the observed distribution. In the graphical representation, the same continuous theoretical distribution may be used for all observed distributions by choosing the scale of the ordinate properly. The length chosen for representing one observation in the initial distribution will represent h observations for the h distributions obtained by the combination of h cells.

The different observed distributions corresponding to the same observations and to the same theory will give different values of

$$\chi^2 = \sum_{i=1}^{k} \frac{(a_i - Np_i)^2}{Np_i}. \tag{6}$$

The expected contents of the first and last cell are

$$Np_1 = NF(y_0 + \Delta y), \tag{7}$$

$$Np_k = N(1 - F(y_0 + (k-1)\Delta y)). \tag{8}$$

Since the total expected frequency must be equal to the number of observations

$$\sum_{i=1}^{k} Np_i = \sum_{i=1}^{k} a_i, \tag{9}$$

formula (6) may be written

$$\chi^2 = \sum_{i=1}^{k} \frac{a_i^2}{Np_i} - N. \tag{10}$$

This formula, being simpler than (6), will be used in the numerical example.

An upper limit for χ^2 is furnished by the case that one cell j contains all observations Then

$$a_j = N; \qquad a_i = 0 \quad \text{for} \quad i \neq j,$$

whence from (10)

$$0 \leqq \chi^2 \leqq \frac{N}{p_j} - N. \tag{11}$$

The upper limit depends again upon the intervals and the starting point of the classification. If the probability for an observation to be contained in the cell j is small, the upper limit is large.

The exact distribution of χ^2 has not yet been established. To obtain an approximation, it is assumed that a binominal distribution may be replaced by a normal distribution. As this does not hold for cells with a small expected frequency, the contents of such cells must be combined. This prescription, which is also valid for a discontinuous variate, constitutes a third arbitrary action in the calculation of χ^2. It invalidates the prior postulate that all cells ought to have the same length.

The approximation used for the probability P of obtaining a value of χ^2, equal to or larger than the observed one, is

$$P(\chi^2, \nu) = K \int_{\chi^2}^{\infty} z^{2\,\frac{1}{2}(\nu-2)} e^{-\frac{1}{2}z^2}\, dz^2 \tag{12}$$

where ν is the number of degrees of freedom. Since

$$\frac{\partial P}{\partial \chi^2} < 0; \qquad \frac{\partial P}{\partial \nu} > 0, \tag{13}$$

P diminishes as χ^2 increases, ν being given, but P increases as ν increases, χ^2 being given. By choosing larger cells, the number ν diminishes, and P may remain the same if χ^2 diminishes adequately.

It is easy to see that χ^2 cannot increase as a result of the combination of cells

and will, in general, decrease. Let a_1 and a_2 represent the actual number of observations in two cells that are to be combined. Let Np_1 and Np_2 be the expected numbers. Then, the contribution of the two separate cells to χ^2 minus the contribution of the two combined cells is, by (10)

$$\frac{a_1^2}{Np_1} + \frac{a_2^2}{Np_2} - \frac{a_1^2 + 2a_1a_2 + a_2^2}{N(p_1 + p_2)}.$$

As a_1 and a_2 are positive or zero, the difference is proportional to

$$a_1^2p_2^2 + a_2^2p_1^2 - 2a_1a_2p_1p_2 = (a_1p_2 - a_2p_1)^2 \geqq 0.$$

The equality holds only when $a_1:a_2 = p_1:p_2$. Then, the combination of cells has no influence on χ^2, but it reduces the number of degrees of freedom by one, and diminishes the probability P. In the general case, the combination of cells diminishes χ^2 and diminishes ν at the same time According to (13), the first influence tends to increase the probability P, the second to diminish it. It cannot be stated a priori which influence is stronger.

For a given set of observations, a continuous variate and a given theory, which includes given estimates of the constants, the probability P depends upon three arbitrary actions. If a certain choice of the intervals gives a good fit, it cannot be concluded that a broader classification gives the same or a better fit [4]. For a given interval, P may vary considerably with the starting point. This influence cannot be allowed for by any formula as the number of degrees of freedom does not depend upon the starting point. Finally, the term "small expected numbers" is vague. Different combinations of cells lead to different probabilities. It is generally assumed that these influences remain within reasonable limits and that P does not vary considerably if we change the class length or the starting point. In the following example, we shall show that this opinion is erroneous.

2. Numerical example. The flood discharge of the Mississippi River at Vicksburg for each of the fifty years 1890–1939 will be used to illustrate the extent to which the observed distributions and P vary with the choice of cell length and the starting point. The observed flood discharges x_m measured in 1,000 cubic feet per second are given in Table VI of a previous article [2], and are not repeated here. The expected distribution is given by the theory of largest values which states that the probability $\mathfrak{W}(x)$ of a flood discharge equal to or less than x is

$$\mathfrak{W}(x) = e^{-e^{-\alpha(x-u)}} \tag{14}$$

Values of $\mathfrak{W}(x)$ as a function of the reduced variate

$$y = \alpha(x - u), \tag{15}$$

are given in Table II of the reference first cited.

Calculation of the constants α and u leads to the theoretical value of the flood discharge

$$x = 1201.9 + 266.1y \tag{16}$$

associated with a given probability $F(y) = \mathfrak{W}(x)$.

TABLE I

Observed and theoretical distribution (1) for the interval $\Delta y = .25$; $\Delta x = 66.525$

Variates		Distributions	
Reduced y	Absolute x	Observed a_i	Theoretical Np_i
1	2	3	4
	736.2	1	.5655
$\leqq -1.50$	802.8	1	.959
−1.25	869.3	3	1.775
−1.00	935.8	3	2.720
−.75	1002.3	5	3.5955
−.50	1068.9	1	4.2315
−.25	1135.4	3	4.5475
.00	1201.9	3	4.554
.25	1268.4	3	4.314
.50	1334.9	6	3.914
.75	1401.5	6	3.434
1.00	1468.0	4	2.934
1.25	1534.6	2	2.4565
1.50	1601.1	0	2.0235
1.75	1667.6	2	1.647
2.00	1734.1	0	1.3270
2.25	1800.6	2	1.0615
2.50	1867.2	2	.844
2.75	1933.7	0	.668
3.00	2000.2	2	.527
3.25	2066.7	0	.414
3.50	2133.3	0	.325
3.75	2199.8	0	.255
4.00	2266.3	0	.1995
$\geqq 4.25$	2332.8	1	.708
		50	50.000

The first observed distribution presented in Table I is obtained by letting $\Delta y = .25$; $\Delta x = 66.525$ and $y_0 = -1.75$. The expected number of observations for the first and last cell are $50F(-1.5)$ and $50\ (1 - F\ (4.25))$ respectively.

The expected frequencies (formula 4) for the other cells

$$np_s = 50\,[F(y + .25) - F(y)],$$

were obtained by successive substraction of two consecutive figures given in column 2, Table II [2]. The theoretical and the observed distribution are plotted in figure 1. The observed distribution given in Table I is very irregular.

Evidently, the intervals are too small. Therefore, we construct the observed and theoretical distributions (2) and (3) for cells which are two times larger.

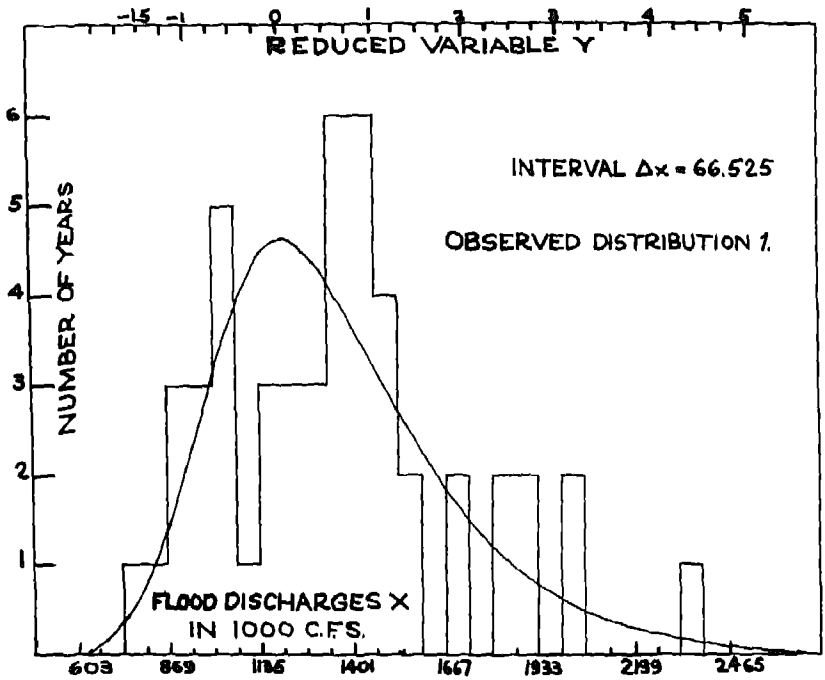


FIG. 1

The first cell in distribution (2) is obtained from distribution (1) by combining the first cell of (1) with the empty one before it; the second cell is obtained by combining the second and third cells of (1); and so on.

Distribution No. 3 is obtained by combining the first two cells of distribution No. 1, then the third and fourth, and so on. The observed distributions 2 and 3 and the theoretical distribution are plotted in figure 2. The scale of the ordinate is $\frac{1}{2}$ of the scale in figure 1. In the same way, the three observed distributions (4), (5), (6) for the interval $\Delta y = \frac{1}{4}$, $\Delta x = 199.57$ are obtained by combining either two void cells with the first cell of Table I, or one void cell with the first and second cell of Table III, or the first three cells of Table I (see fig. 3).

Finally, the four observed distributions (7), (8), (9), (10) for the interval

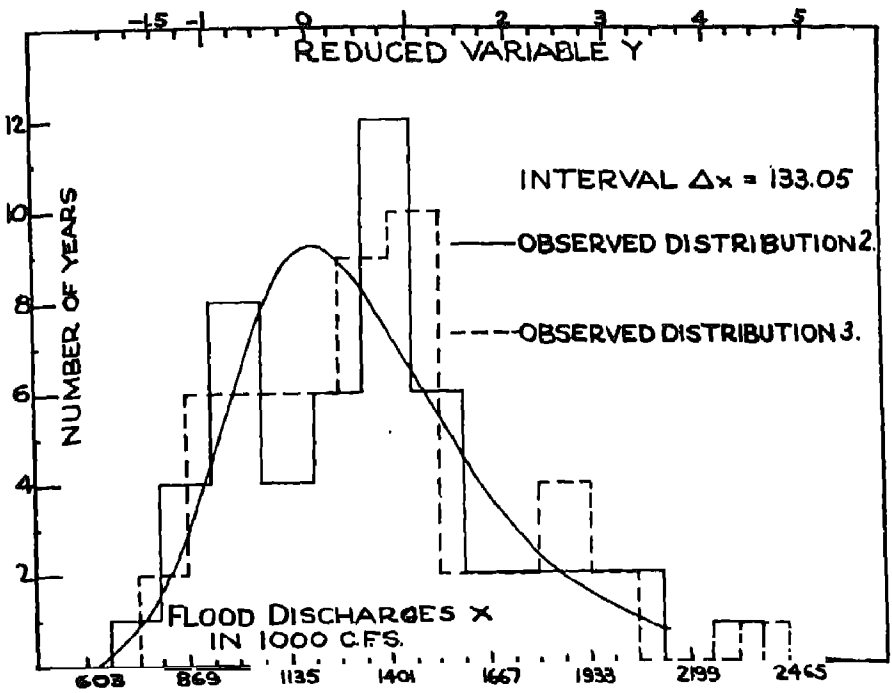

REDUCED VARIABLE Y
-1.5
-1
0
2
3
4
5
NUMBER OF YEARS
12
10
8
6
4
2
INTERVAL Δx = 133.05
OBSERVED DISTRIBUTION 2.
OBSERVED DISTRIBUTION 3.
FLOOD DISCHARGES X
IN 1000 C.F.S.
603
869
1135
1401
1667
1933
2199
2465

FIG. 2

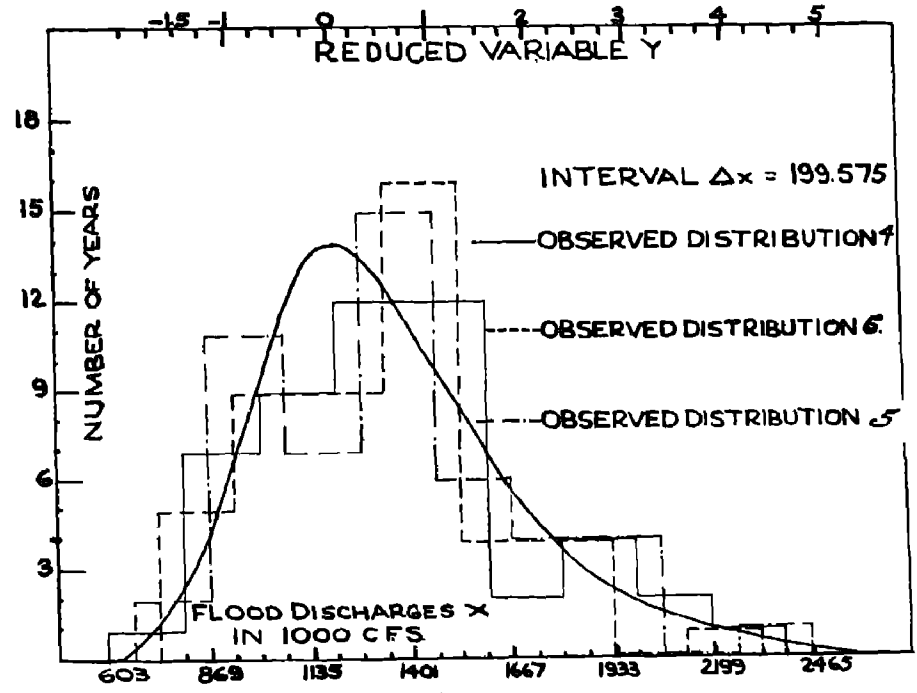

REDUCED VARIABLE Y
-1.5
-1
0
2
3
4
5
NUMBER OF YEARS
18
15
12
9
6
3
INTERVAL Δx = 199.575
OBSERVED DISTRIBUTION 4
OBSERVED DISTRIBUTION 6.
OBSERVED DISTRIBUTION 5
FLOOD DISCHARGES X
IN 1000 C.F.S.
603
869
1135
1401
1667
1933
2199
2465

FIG. 3

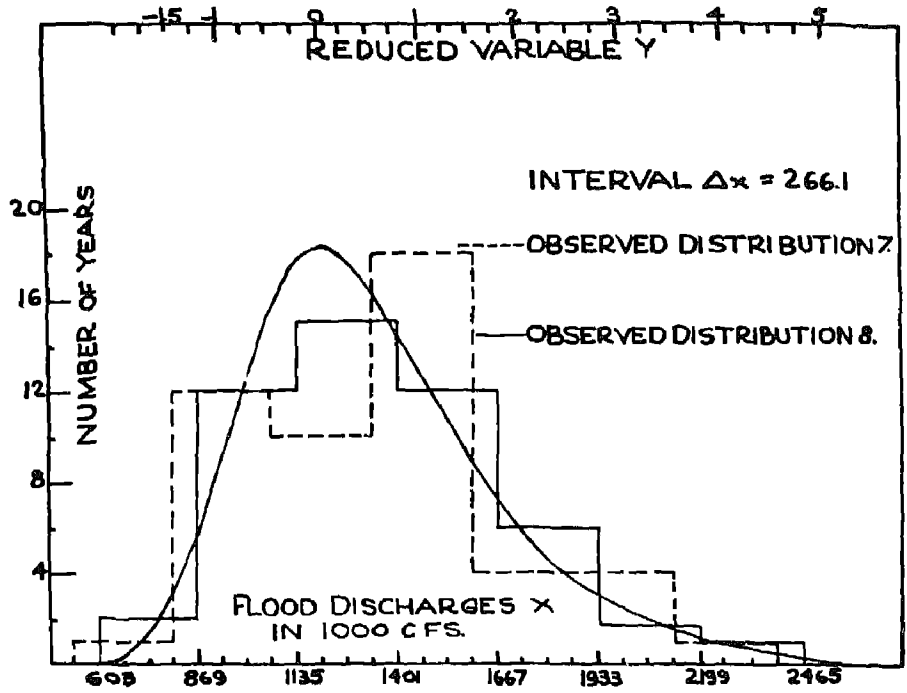

REDUCED VARIABLE Y
-1.5
-1
0
1
2
3
4
5
INTERVAL Δx = 266.1
OBSERVED DISTRIBUTION 7
OBSERVED DISTRIBUTION 8.
NUMBER OF YEARS
20
16
12
8
4
FLOOD DISCHARGES X
IN 1000 C.F.S.
603
869
1135
1401
1667
1933
2199
2465

FIG. 4

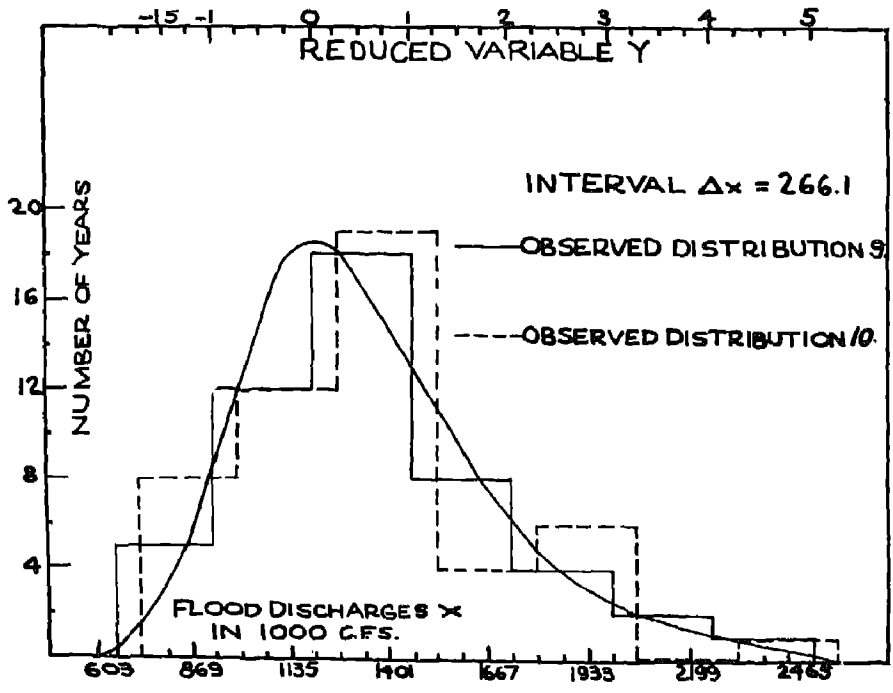

REDUCED VARIABLE Y
-1.5
-1
0
1
2
3
4
5
INTERVAL Δx = 266.1
OBSERVED DISTRIBUTION 9.
OBSERVED DISTRIBUTION 10.
NUMBER OF YEARS
20
16
12
8
4
FLOOD DISCHARGES X
IN 1000 C.F.S.
603
869
1135
1401
1667
1933
2199
2465

FIG. 5

$\Delta y = 1$; $\Delta x = 266.1$ are compared with the theoretical distribution in figures 4 and 5. The four distributions 7–10 differ considerably. Distributions 8 and 9 indicate that the agreement between theory and observations is good, distribution 7 and 10 indicate that the fit is bad. The χ^2 method must give the same contradictory results.

TABLE II

Four values of $P(\chi^2)$ for the same observations and the same theory

1	2	3	4	5	6	7	8	9	10
Midpoints	Observed Distributions, a_i				Theoretical Distributions, Np_i	Components of $\chi^2 + N$			
	(7)	(10)	(9)	(8)					
803			5		3.2995			7.577	
869		8			6.0195		10.632		
936	13				9.6150	17.577			
1002				14	13.8465				14.155
1069			12		15.0945			9.540	
1135		12			16.9285		8.506		
1202	10				17.6470	5.667			
1268				15	17.3295				12.984
1335			18		16.2160			19.980	
1401		19			14.5960		24.733		
1468	18				12.7385	25.435			
1534				12	10.8480				13.274
1601			8		9.0610			7.063	
1667		4			7.4540		2.146		
1734	4				6.0590	2.641			
1800				6	4.8795				7.378
1867			7		6.3290			7.742	
1933		7			5.0020		9.796		
2000	5				2.9405	6.344			
2066				3	3.0965				2.907
N .	50	50	50	50	200.0000	$\chi^2 + N = 57.664$	55.813	51.902	50.698
ν. .	2	2	2	2	P .	.023	.057	.399	.705

The details for the calculations of χ^2 are given in Table II. The numbers of column 1 are the midpoints of the cells. To save space, the four theoretical distributions obtained from Table I, col. 4 are written in the same column (6) directly opposite the corresponding observed distributions given in columns 2 to 5. Through formula (10) we calculate the components of $\chi^2 + N$ (cols. 7 to 10). Although the four distributions differ only with respect to the beginning

of the first cell, the value of P for the observed distribution number (8) is more than thirty times the value of P for the observed distribution number (7). In view of the fact that these values of P are calculated for a fixed set of observations, for the same theory, the same constants, and the same number of degrees of freedom, the differences found are surprising.

3. The probability integral transformation. This example shows that the probability P may vary with the starting point in such a way that no conclusion about the acceptance or rejection of a hypothesis can be obtained from the usual χ^2 method. The three arbitrary steps described above may be avoided if we choose cells of equal probability instead of cells of equal length. The required intervals are obtained from the probability integral transformation, due to Karl Pearson [6]. Let $w(x)$ be a distribution of a continuous variate x, let $y = W(x)$ be the transformed variate, then the distribution $p(y)$ of the variate y is

$$p(y) = 1. \tag{17}$$

In other words: The probabilities $W(x)$ are uniformly distributed. If a distribution $w(x)$ has been chosen for a given set of observations x_m, we can control this theory by investigating whether the "observations" $W(x_m)$, i e, the theoretical cumulative frequencies of the observed values are uniformly distributed. Thus, the comparison of the observed distributions with any continuous theoretical distribution is reduced to the comparison of an "observed" with a theoretical uniform distribution. To a given set of observations and a given theory there is one, and only one, "observed" distribution. If we introduce within $w(x)$ another set of constants, or choose instead of $w(x)$ another theory $\varphi(x)$, we obtain, of course, other "observed" values [1].

The goodness of fit between this theory and these "observations" may be measured by the χ^2 method. We divide the interval zero to N, which contains the N "observed" numbers $NW(x_m)$ into k cells of equal length, and enumerate the "observed" points $NW(x_m)$ contained in each cell. The starting point of the classification is always zero. The expected number of observations for each cell is always N/k. If we choose k sufficiently small, the necessity for combining cells is eliminated. We have to choose k in such a way that the conditions, under which formula (12) holds, are fulfilled. The question of the best choice for the number of cells has been studied by Wald and Mann [3]. Their solution is valid for small levels of significance and for large numbers of observations.

4. Conclusion. The usual χ^2 test is unreliable for a continuous variate as it involves three arbitrary decisions. From the same observations, the same theory, and the same constants different statisticians, equally well trained and equally careful, may obtain different probabilities P, and may proclaim any one of these results as final. Therefore, the usual χ^2 method does not lead to a decision whether a hypothesis has to be rejected or not. Such a decision is possible if we use the probability integral transformation. Unfortunately, the question

of the best choice of the cells for small numbers of observations and large levels of significance is not yet solved.

REFERENCES

[1] E. J GUMBEL, "Simple tests for given hypotheses," *Biometrika*, Vol 32, Parts 3 and 4 (1942), pp. 317–333.

[2] E J GUMBEL, "The return period of flood flows," *Annals of Math. Stat.*, Vol 12 (1941), pp 163–190.

[3] H. B MANN and A. WALD, "On the choice of the number of class intervals in the application of the chi square test," *Annals of Math Stat.*, Vol 13 (1942), p 306–317.

[4] J. NEYMAN and E. S. PEARSON, "Further notes on the χ^2 distribution," *Biometrika*, Vol. 22, Parts 3 and 4 (1931), pp. 298–305

[5] KARL PEARSON, "On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random," *Biometrika*, Vol. 25, Parts III and IV (1933), pp. 379–410.

A SAMPLING INSPECTION PLAN FOR CONTINUOUS PRODUCTION[1]

By H. F. Dodge

Bell Telephone Laboratories, New York

I. Introduction

1. Purpose. This paper presents a plan of sampling inspection for a product consisting of individual units (parts, subassemblies, finished articles, etc.) manufactured in quantity by an essentially continuous process.

The plan, applicable only to characteristics subject to nondestructive inspection on a Go-NoGo basis, is intended primarily for use in process inspection of parts or final inspection of finished articles within a manufacturing plant, where it is desired to have assurance that the percentage of defective units in accepted product will be held down to some prescribed low figure. It differs from others which have been published[2,3] in that it presumes a *continuous flow* of *consecutive articles* or *consecutive lots* of articles offered to the inspector for acceptance in the order of their production. It is accordingly of particular interest for products manufactured by conveyor or other straight line continuous processes.

In operation, the plan provides a corrective inspection, serving as a partial screen for defective[4] units. Normally, a chosen percentage or fraction f of the units are inspected, but when a defective unit is disclosed by the inspection it is required that an additional number of units be inspected, the additional number depending on how many more defective units are found. The result of such inspections is to remove some of the defective units, and the poorer the quality submitted to the inspector, as measured in terms of per cent defective, the greater will be the corrective or screening effect. The object of the plan is the same as that incorporated in some of the sampling tables already published[5], namely, to establish a limiting value of "average outgoing quality" expressed in per cent

---

[1] Presented at the Joint Meeting of the American Society of Mechanical Engineers and the Institute of Mathematical Statistics, May 29, 1943, by H. F. Dodge, Quality Results Engineer, Bell Telephone Laboratories, New York

[2] H. F. Dodge and H. G. Romig, "Single Sampling and Double Sampling Inspection Tables", *Bell Sys Tech Jour*, Vol XX (1941) pp. 1–61. An unpublished paper by Prof Walter Bartky (developed when he was associated with the Western Electric Co., 1927) provides a continuous multiple sampling plan involving two factors—f, as used here, and i, the number of units in a "compensating sample" required to be inspected for each defective unit found.

[3] Lt. R. J Saunders, "Standardized Inspection", *Army Ordnance*, Vol XXIV (1943) pp 290–292; G Rupert Gause, "Quality Through Inspection", *Army Ordnance*, Vol. XXIV (1943) pp 117–120

[4] A unit of product that fails to meet the requirement for a characteristic is classed as nonconforming with respect to that characteristic, and for convenience is referred to as "defective" Thus, a deviation from a specified requirement or from accepted standards of good workmanship is termed a "defect"

[5] H. F. Dodge and H. G Romig, loc cit.

defective which will not be exceeded no matter what quality is submitted to the inspector. This limiting value of per cent defective is termed the "average outgoing quality limit (AOQL)".

The theoretical solution treats the case of inspecting a continuous flow of individual units and is based on the distribution of *random-order* spacing of defective units in product whose quality is statistically controlled.[6] Part III of the paper extends the application of the method to a continuous flow of individual lots or sub-lots of articles.

II. Inspection of a Flow of Individual Units

2. Inspection of one characteristic. Consider first the inspection of a flow of individual units, offered consecutively in the order of their production. Assume that inspection is to be made for only one quality characteristic, so that interest will be centered on one kind of defect. Subsequently (Section 13), consideration will be given to the procedures when inspection is made simultaneously for several kinds of defects.

3. Procedure A. The procedure is as follows:

(a) At the outset, inspect 100% of the units consecutively as produced and continue such inspection until i units in succession are found clear of defects.

(b) When i units in succession are found clear of defects, discontinue 100% inspection, and inspect only a fraction f of the units, selecting individual sample units one at a time from the flow of product, in such a manner as to assure an unbiased sample.

(c) If a sample unit is found defective, revert immediately to a 100% inspection of succeeding units and continue until again i units in succession are found clear of defects, as in paragraph (a).

(d) Correct or replace with good units, all defective units found.

4. Protection provided by the plan. The inspection plan is defined by the two constants, f and i, which can be altered at will. For given values of f, i, and p (incoming fraction defective), there will result for product of statistically controlled quality a definite average outgoing fraction defective (average outgoing quality, AOQ). For given values of f and i, the AOQ will have a maximum for some particular fraction defective p_1 of incoming quality. As noted above, this maximum is referred to as the average outgoing quality limit (AOQL). For all other values of incoming fraction defective p greater or less than p_1, the AOQ will be less than AOQL Many combinations of f and i will result in the same AOQL

The protection offered by the plan discussed here can thus be expressed in terms of the AOQL, in per cent defective.

---

[6] "Statistical control" as defined in the literature, see W A Shewhart, *Statistical Method from the Viewpoint of Quality Control*, The Graduate School, U. S. Dept. of Agriculture, 1939.

5. Theoretical framework. We are concerned with the spacing between defective units when the individual units are arrayed in the order of their production, as shown in Fig. 1. If the manufacturing process is statistically controlled so that the probability of producing a defective unit is constant and equal to p, then defective units will have an order spacing of a random character which is expressible in terms of certain probability laws. Product turned out by such a process will be referred to as having a process average fraction defective p. The "event" of particular interest is a "terminal-defect sequence" of $i + 1$ successive units following the observance of a defect, comprising a succession of i nondefective units followed by a defective unit, as shown in Fig. 1. The totality of all possible such sequences, where i varies from 0 to ∞, constitutes the universe of events under consideration.

Each such sequence of $i + 1$ units, comprising i successive nondefective units followed by a defective one, has a definite probability of occurrence, for a process average fraction defective, p. The complete set of such probabilities for all possible sequences, having respectively $i = 0, 1, 2, 3, \cdots \infty$, defines a probability distribution[7] of random-order spacing of defects in uniform product. This is

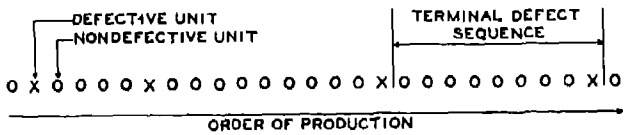


FIG. 1. Spacing of defective units

shown in the table below in which 0 represents a nondefective unit, X represents a defective one, p is the fraction defective, and $q = 1 - p$.

Sequence	*Defect Spacing (No. of units in sequence)*	*No. of Non-defective Units before Finding the Next Defect*	*Proba-bility of Occur-rence*	*No. of Term in the Power Series*
X	1	0	p	1st
$0X$	2	1	pq	2nd
$00X$	3	2	pq^2	3rd
$000X$	4	3	pq^3	4th
$0000X$	5	4	pq^4	5th
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
$000\cdots0X$	$i + 1$	i	pq^i	$(i + 1)st$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

[7] Romanovsky, V., "Due Nuovi Criteri di Controllo Sull 'andamento Casuale di Una Successione di Valori", *Giornale dell 'Instituto Italiano degli Attuari* (1932) discusses this

These probabilities are the successive terms in the infinite power series

$$p + pq + pq^2 + pq^3 + \cdots$$

$$\text{or}, \quad p(1 + q + q^2 + q^3 + \cdots). \tag{1}$$

The sum of this series is $p\left(\frac{1}{1-q}\right) = 1$, i.e., the total probability for all possible sequences is unity (as it should be).

The sum of the first $i + 1$ terms of the series is the probability of occurrence of a "terminal-defect sequence" (defect spacing) of $i + 1$ units or less. The sum of the first i terms is the probability, P_1, of failing to find the next i units clear of defects, which is

$$P_1 = \sum_{j=0}^{j=i-1} pq^j = 1 - q^i. \tag{2}$$

In turn, the sum of all terms beyond the ith term is the probability of finding 0 defects in the next i units, which is

$$Q_1 = 1 - P_1 = q^i. \tag{3}$$

These results and the power series (1) enter into subsequent portions of the discussion. The curves of Fig. 2 give values of $1 - q^i$.

6. Average outgoing quality. Suppose a plan is selected, choosing specific values of f and i.

For given values of i and p, there will be an expected average number of units, u, inspected following the finding of a defect. Likewise, for given values of f and p there will be an expected average number of units, v, that will be passed under the sampling procedure before a defect is found. The latter average number includes the sampling units actually inspected as well as the uninspected units produced between successive sample units.

The average fraction of the total product units inspected in the long run is

$$F = \frac{u + fv}{u + v}. \tag{4}$$

It is now assumed for purposes of solution that the inspection operation itself never overlooks a defect and that all defective units found during the inspection of f and i will be corrected or replaced by good units.[8]

---

probability distribution of spacing of events, referring to the spacing as the "length of a partial series" Our term "terminal-defect sequence" has the same significance as his term "partial series". See also P. S Olmstead, "Note on theoretical and observed distributions of repetitive occurrences", *Annals of Math Stat* Vol XI (1940) pp 363–366; A M. Mood, "The distribution theory of runs", *Annals of Math. Stat* , Vol XI (1940) pp. 367–392

[8] The assumption that the inspection operation is perfect cannot be made without reservation Machine inspection devices have their margins of error Also, inspection fatigue prevents 100% manual and visual inspections from insuring perfection, particularly if such inspections continue over a considerable period of time But the efficiency of the latter

As a result of the screening effect of the inspection, the average outgoing quality, AOQ, designated p_A, is related as follows to the incoming quality p:

$$p_A = p(1 - F) = p\left(1 - \frac{u + fv}{u + v}\right). \tag{5}$$

7. Determination of u. The average number of units, u, inspected on a 100% inspection basis following the finding of a defect is a function of i and p, and may be determined from a considration of two power series, one limited and the other infinite.

Once the 100% inspection starts, there are several things that can happen before i units are found clear of defects The first i may be found clear; or 1, 2, 3, or more defects may be found before finally a run of i units is found clear.

One of the quantities to be determined is the *average number of units* inspected in a "failure sequence," that is, one terminating in a defect and comprising i or less units. This average number, designated as h, is the average of the distribution made up of the first i terms of the power series (1). The average is

$$h = \frac{p}{1 - q^i}(1 + 2q + 3q^2 + 4q^3 + \cdots + iq^{i-1}), \tag{6}$$

where the denominator is the sum of the probabilities for the first i terms. This may be evaluated as follows:

$$\begin{aligned} h &= \frac{p}{1 - q^i}\frac{d}{dq}(1 + q + q^2 + q^3 + \cdots + q^i) \\ &= \frac{p}{1 - q^i}\frac{d}{dq}\left[\frac{1 - q^{i+1}}{1 - q}\right] \\ &= \frac{1}{p(1 - q^i)}[1 - q^i(1 + pi)]. \end{aligned} \tag{7}$$

Note that if pi is small compared with unity, h is approximately $1/p$.

The next step is to determine the *average number of failure sequences* that will be encountered before finding i units clear of defects. This average number, designated as G, may be found from the probability distribution of all possible numbers of failure sequences, expressed by the infinite series

$$Q_1(1 + P_1 + P_1^2 + P_1^3 + \cdots) \tag{8}$$

where P_1 is given by equation (2), $Q_1 = 1 - P_1$, as given by equation (3), and the successive terms are the probabilities of occurrence of 0, 1, 2, 3, etc. failure

---

inspections is generally higher when an interest incentive is provided as is usually the case in sampling inspection plans where the extent of such inspections hinges on their findings.

The solution given assumes correction or replacement of defective units. Where it is expedient to reject such units and not replace them, equations (19) to (22) inclusive, should be modified by replacing i by $i - 1$.

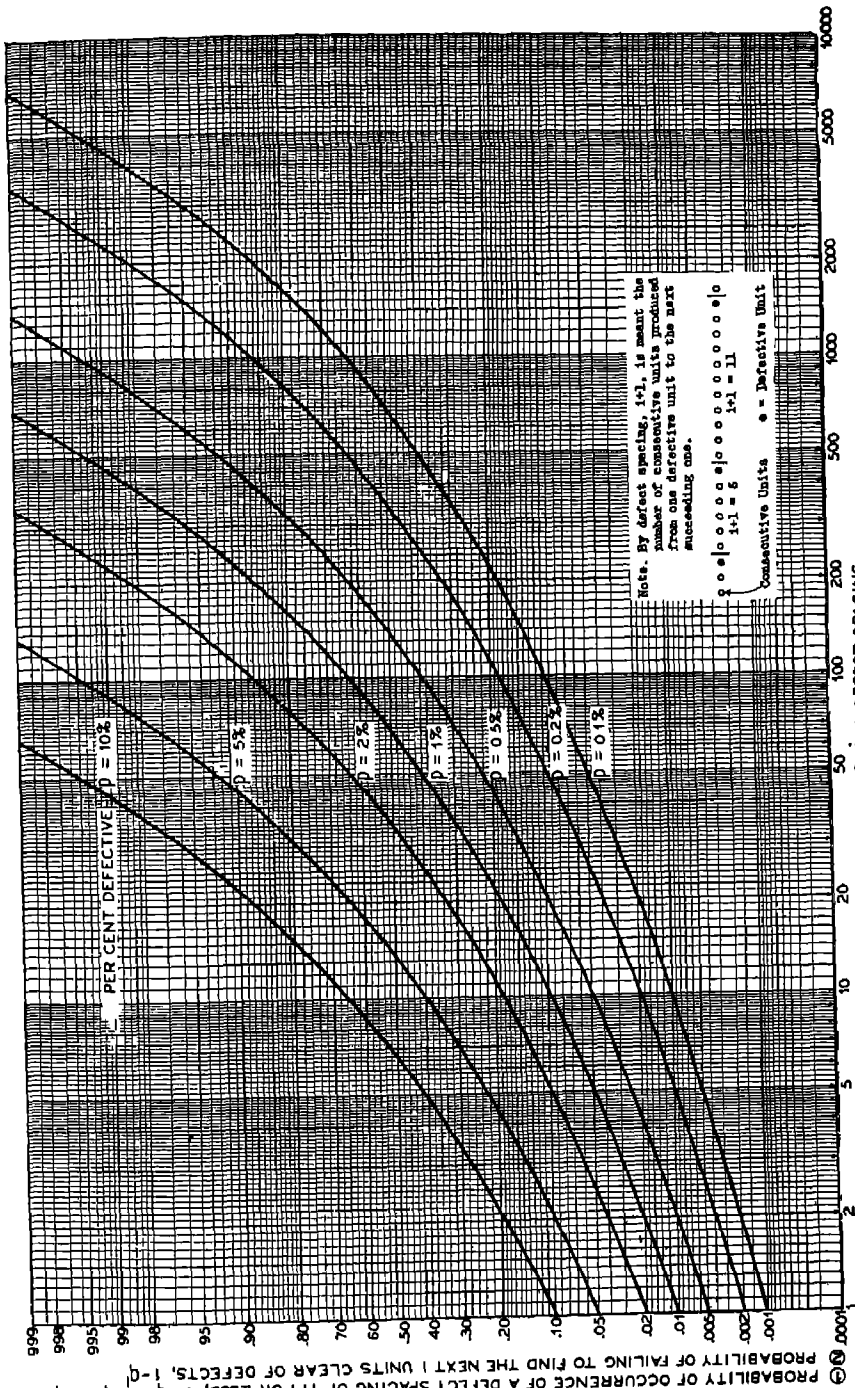

PROBABILITY OF OCCURRENCE OF A DEFECT SPACING OF i+1 OR LESS, $1-q^{i+1}$ ($q = 1-p$)
PROBABILITY OF FAILING TO FIND THE NEXT i UNITS CLEAR OF DEFECTS, $1-q^i$
PER CENT DEFECTIVE
p = 10%
p = 5%
p = 2%
p = 1%
p = 0.5%
p = 0.2%
p = 0.1%
Note. By defect spacing, i+1, is meant the number of consecutive units produced from one defective unit to the next succeeding one.
Consecutive Units
i+1 = 5
i+1 = 11
o = Defective Unit
① i+1, DEFECT SPACING
② i, NUMBER OF UNITS

FIG 2 Curves defining distribution of random order spacing of defects in uniform product

sequences before finding i units clear of defects. The average number of failure sequences, G, is given by the sum of the infinite series

$$G = Q_1(0 + 1P_1 + 2P_1^2 + 3P_1^3 + \cdots)$$

$$= Q_1P_1(1 + 2P_1 + 3P_1^2 + 4P_1^3 + \cdots). \tag{9}$$

Summing the series, we have

$$G = Q_1P_1\frac{1}{(1 - P_1)^2} = \frac{P_1}{Q_1} = \frac{1 - q^i}{q^i} \tag{10}$$

Now u, the average number of pieces inspected following the finding of a defect, is made up of a number of failure sequences followed by a run of i units clear of defects. Using the average values of G and h just found, we have

$$u = Gh + i = \frac{1 - q^i}{pq^i}. \tag{11}$$

8. Determination of v. The average number of units, v, that will be passed in a period of sampling inspection will be $1/f$ times the average number of individual sample units inspected in such periods. Here again the solution will depend on the random order spacing of defects in uniform product. Whether the individual units selected during the sampling inspection procedure are selected by a random spacing device, or by any other means which will prevent known bias in the sample, we may assume that defects will be found to occur in accordance with the distribution of random order spacing defined by the terms of the series given in (1). The average number of sample units inspected in a period of sampling inspection will thus be the average defect spacing for product having fraction defective, p, which is given by the infinite series.

$$H = p(1 + 2q + 3q^2 + 4q^3 + \cdots). \tag{12}$$

Summing the series, we have

$$H = \frac{p}{(1 - q)^2} = \frac{1}{p}, \tag{13}$$

and the value of v is found to be

$$v = \frac{H}{f} = \frac{1}{fp}. \tag{14}$$

9. Determination of f and i for a given value of AOQL. From the considerations given above, the average fraction of the product inspected, F, and the value of average outgoing quality, p_A, can be determined for any given values of p, f, and i. Substituting in (5), the values of u and v given in (11) and (14), we have

$$p_A = p\left[1 - \frac{f}{f + (1 - f)(1 - p)^i}\right]. \tag{15}$$

The average outgoing quality limit, AOQL, (p_L) is the maximum value of p_A that will result for any given values of f and i, considering all possible values of p in the submitted product. The value of p for which this maximum value of p_A occurs is designated by p_1, hence

$$p_L = p_1\left[1 - \frac{f}{f + (1-f)(1-p)^i}\right]. \tag{16}$$

The value of p_1 for which $p_A = p_L$ is determined by differentiating (15) with respect to p, equating to 0, and solving for p, that is

$$\frac{dp_A}{dp} = 1 - \frac{f^2 + f(1-f)(1-p)^i + pfi(1-f)(1-p)^{i-1}}{[f + (1-f)(1-p)^i]^2}. \tag{17}$$

Simplifying, and using the designation p_1 for the maximizing value of p, gives

$$(i+1)p_1 - 1 = \frac{1-f}{f}(1-p_1)^{i+1}, \quad \text{or}$$

$$(1-p_1)^i = \frac{f[(i+1)p_1 - 1]}{(1-f)(1-p_1)}. \tag{18}$$

Substituting in (16) this value of $(1-p_1)^i$, we have

$$p_L = \frac{(i+1)p_1 - 1}{i}, \quad \text{hence} \tag{19}$$

$$p_1 = \frac{1 + ip_L}{i+1}. \tag{20}$$

From (18) and (19), we have

$$p_L = \frac{1-f}{fi}(1-p_1)^{i+1}, \quad \text{hence} \tag{21}$$

$$f = \frac{(1-p_1)^{i+1}}{ip_L + (1-p_1)^{i+1}}. \tag{22}$$

The curves given in Fig. 3 were calculated by choosing values of i for given values of AOQL (p_L) and calculating p_1 from equation (20) and f from equation (22). Thus for a given AOQL value, an i value may be found for a chosen f value and vice versa. It will be noted that for a given value of f, i varies inversely with the AOQL value, to a close degree of approximation.

10. Operating characteristics of the plan. Figs. 4(a) and 4(b) give a picture of the operating characteristics of the general plan as f and i are varied. They indicate for example that for a moderate range of f values the factor i has a stronger influence than f in determining the discrimination that the method affords between high and low levels of incoming per cent defective. For the values of f and i shown, Fig. 4(b) indicates just what level of incoming per cent

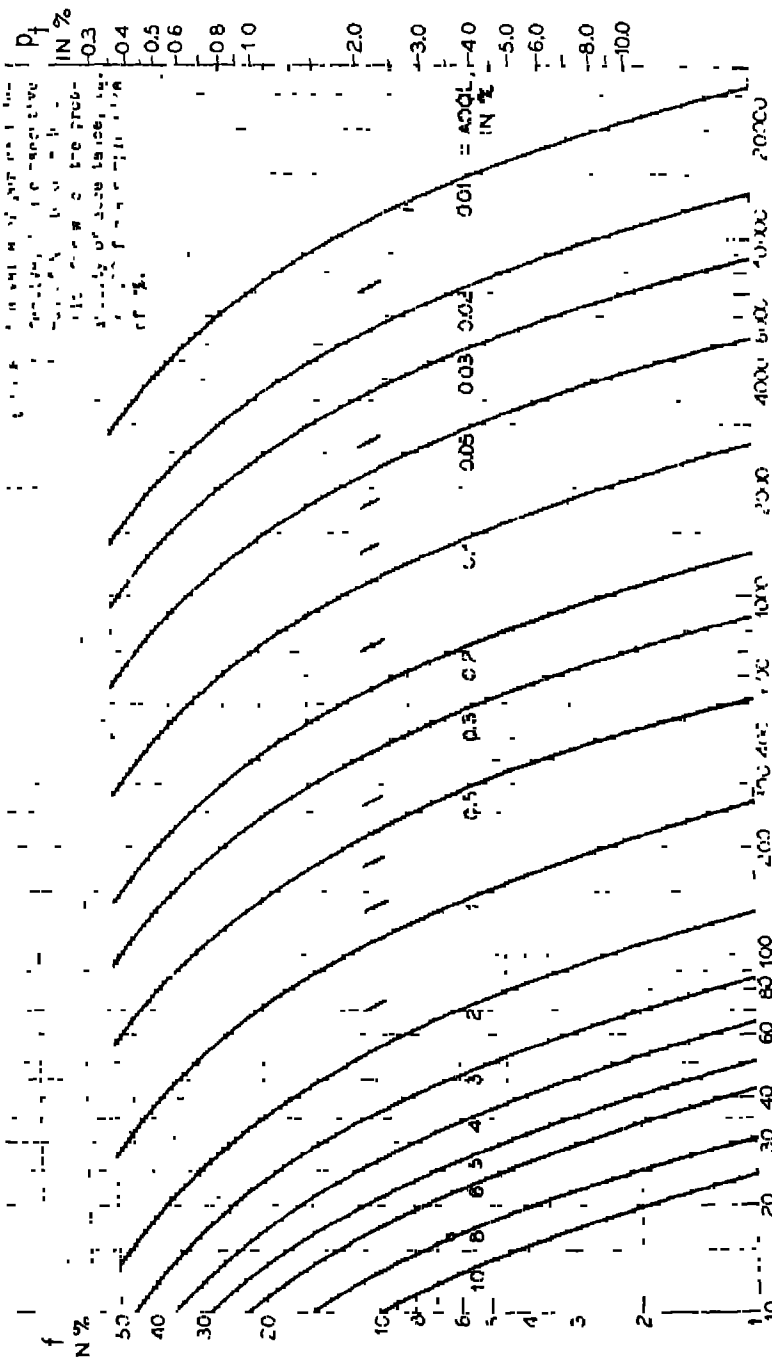


FIG 3 Curves for determining values of f and n for a given value of AOQL

defective would force a correction of the manufacturing process, if the percentage of total production that would be accepted on a sampling basis falls below a critical value—often, a value of the order of 80% to 90%.

Fig. 5 gives a comparison of the characteristics of several plans having the same AOQL value, 1%. It indiates for example that when the normal level of incoming per cent defective is well below the AOQL, the AOQL value can be assured with less inspection by choosing f small and i large But since, for a given AOQL value, the average amount of inspection approaches a minimum as f approaches 0, factors other than the minimum amount of inspection have a

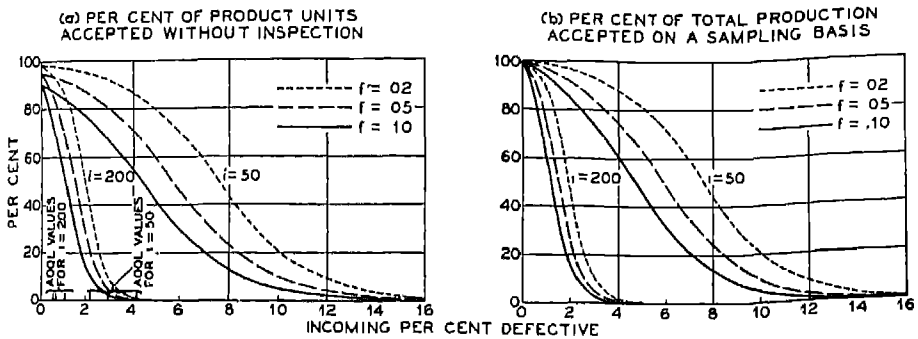

(a) PER CENT OF PRODUCT UNITS ACCEPTED WITHOUT INSPECTION
(b) PER CENT OF TOTAL PRODUCTION ACCEPTED ON A SAMPLING BASIS
PER CENT
f = 02
f = 05
f = 10
f = .10
i = 200
i = 50
AOQL VALUES FOR i = 200
AOQL VALUES FOR i = 50
INCOMING PER CENT DEFECTIVE

FIG 4. Curves showing effect of f and i on operating characteristics of plan

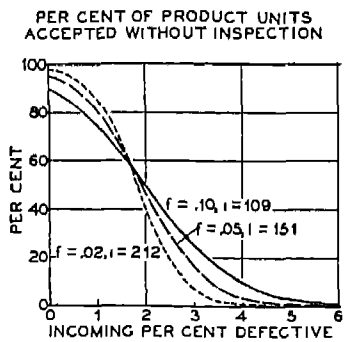


FIG. 5 Characteristics of three plans having the same AOQL of one per cent

more important influence on the choice of the most advantageous combination of f and i values for a given set of circumstances. For example, when the inspector is located at the end of the production line, it may be desirable to use a value of i not greater than some small multiple of the number of product units on the line at any one time. Or again, the value of f is often influenced by the normal work loads of the inspector and the operators on the line. Protection against "spotty" quality, such as may arise from temporary irregularities in workmanship or materials, should receive special consideration in connection with the choice of f.

11. Protection against spotty quality. The p_t scale at the right of Fig 3 provides a guide concerning the protection afforded against spotty quality in a continuous run of product. The value of p_t is the per cent defective in a run of 1000 consecutive product units, for which the probability of acceptance by sample is 0.10 for a percentage sample equal to the corresponding f value shown on the chart.

This scale indicates that the protection against spotty quality falls off very rapidly with f and that the protection, considering runs of product of 1000 consecutive units each, becomes quite poor if f is less than 2%.

12 Effect of selecting group samples rather than one unit at a time. The above development assumes selection of individual sample units one at a time from the flow of product and immediate examination of a unit to determine whether or not it is defective. Deviations from this procedure will in general result in giving values of AOQL higher than those shown in Fig. 3.

For example, the actual AOQL may be higher than the theoretical value (a) if the inspector delays looking at the individual units immediately when they are withdrawn from the line, or (b) if he selects a group of units at one time from the production line. The effect of either of these two deviations, both constituting a delay, may be quite large if i is small, or if large group samples are taken.

Although the modification of the theoretical AOQL value resulting from the selection of group samples has not been thoroughly explored, this should not be excessive,

(a) if group samples of $n = 10$ or less are drawn from the line, and

(b) if $i = 50$ or more,

provided there is no delay in examining the group samples drawn from the line.

It should be noted however, that the effect of these delay factors on the AOQL may be compensated for in part if, when a defect is found, the 100% inspection includes some of the units that have already passed the inspection point.

Where appreciable delays are unavoidable, an alternative is to withhold from acceptance a stipulated number of units pending the examination of the sample units that have been selected to represent this quantity of product. Such a procedure provides in effect a lot acceptance plan, the treatment for which is covered in Part III.

13. Administration of inspection operations. The inspection plan is most effective in practice if it is administered in such a way as to provide an incentive to clear up causes of trouble promptly. Such an incentive may be had by imposing a penalty on the operating or manufacturing department when defects are encountered. Normally, no such penalty is imposed if both the sampling inspection and the 100% inspection are performed by the same person or group of persons and the two costs merged; the inspector then merely serves as an

agency for screening defects when quality goes bad It is accordingly recommended that the sampling inspection and the 100% inspection operations be treated as two separate functions

With this in mind, the inspection work can be performed by two different inspectors, designated inspector C and inspector M. Inspector C may be considered as the *consumer's* representative in that his work is performed as a function independent of the manufacturing group. The term "consumer" is used in the general sense of the recipient of the product after the inspection has been completed. Inspector M is responsible to the Manufacturing Department and the cost of his work is borne by that Department. His work must however be subject to the surveillance and approval of inspector C.

The following method of administering the inspection plan can then be used:

(a) Inspector C inspects the required fraction f. So long as no defects are found, product is considered acceptable and is passed.

(b) When inspector C finds a defect, he
   1. continues inspecting the fraction f,
   2. places some identification on the succeeding flow of product to indicate nonacceptance (or diverts it from the regular production line if the design of the line permits), such designation to apply until clearance is obtained in accordance with paragraph (c), and
   3 calls inspector M to inspect the succeeding flow of product in accordance with paragraph (c)

(c) Inspector M (one or more inspectors as needed) inspects all succeeding units, except those inspected by inspector C in the fraction f, until the required number of units, i, are found clear of defects. Inspector M reports immediately to Inspector C all defects found in the course of his 100% inspection and notifies him when a run of i units has been found clear of defects.

(d) When notified that a run of i units has been found clear of defects, inspector C, if satisfied with the work of inspector M, releases inspector M.

(e) To facilitate speedy correction of causes of trouble, inspector C, on finding a defect, should promptly notify the production foreman or other designated authority and furnish the latter with detailed information regarding the character of the defect found.

It will be noted that the above procedure requires calling inspector M whenever inspector C finds a defect. To avoid taking such action on the occurrence of a single defect, the procedure can be modified so that inspector M is called into the picture only when two defects in succession are observed by inspector C. Where this feature is desired, paragraph (b) above may be modified to read as follows:

(b) When inspector C finds a defect, he
   1. proceeds immediately to inspect all succeeding units up to a total of i units, and if no defects are found therein, he again limits his inspection

to the fraction f. If, on the other hand, during the course of inspecting the next i units, inspector C finds a second defect, he immediately discontinues his 100% inspection,

2. places some identification on the succeeding flow of product .. etc.

While this procedure carries the disadvantage of placing a varying work load on inspector C, it is often preferred since a single defect tends to be regarded as an isolated occurrence whereas two defects in quick succession, (like a first and second offense) are normally accepted as sufficient evidence to justify special action.

14. Inspection for several kinds of defects simultaneously. The procedure given above may be applied directly to an inspection covering two or more kinds of defects, provided that the chosen AOQL value applies to all defects collectively and each unit inspected is always inspected for all of the defects under consideration.

It is sometimes desired, however, when a defect of one kind is observed, to confine the 100% inspection to this one kind of defect alone. This requires a modification of the general procedure and the establishment of a separate AOQL for each kind of defect. A similar modification is required for example where the inspection is to cover several kinds of defects, but where the defects are grouped into two or more classes, according to their seriousness, and the defects in each class treated collectively.

The following paragraphs outline for illustrative purposes a procedure for use where the defects under consideration are to be classified into two groups, Major and Minor, and where all Major defects are to be treated collectively and all Minor defects likewise By analogy, the procedure to be followed when each kind of defect is to be treated separately will be obvious. In any event, the fraction f is made the same for all classes or all kinds of defects.

Procedure

Several kinds of defects are grouped into two classes with respect to seriousness; designated Major and Minor.

All defects of the same class (Major or Minor) are treated collectively.

Preliminary

(1) Establish an overall AOQL value for Major defects and an overall AOQL value for Minor defects. Select a suitable value for f, applicable to both Major and Minor defects From Fig. 3 determine a value of i for Major defects, designated i_A, and a value of i for Minor defects, designated i_B.

(2) At the outset, inspect 100% of the units consecutively for both Major and Minor defects until i_{Max} units in succession are found clear of defects ($i_{Max} = i_A$ or i_B, whichever is the larger).

Routine

(3) When i_{Max} units in succession are found clear of defects, discontinue 100% inspection and inspect only a fraction f of the units *for both Major and Minor defects*, selecting individual sample units one at a time from the flow of product.

(4) If a Major (or Minor) defect is observed during sampling inspection, inspect 100% of the succeeding units *only for defects of the class in question* until i_A (or i_B) units in succession are found clear of defects of this class.

(4.1) During the 100% inspection referred to in (4) inspect a portion f *for both Major and Minor defects.*

(4.2) If during the 100% inspection for a particular class of defect (Major or Minor), a defect of the other class is observed on an individual unit of product, start 100% inspection for defects of the new class *only if* the new defect is observed on one of the f units that has been inspected for both Major and Minor defects, and continue such 100% inspection for defects of the new class until i (as determined in (1) for the new class) units in succession are found clear of defects of the new class. Do not take such action, however, if the new defect happens to be observed on one of the non-f units.

(5) When the proper number of successive units are found clear of defects as in paragraph (4) or (4.2), reinstate sampling inspection as in paragraph (3).

From the above it may be appreciated that difficulties of administration are introduced in treating a large number of classes of defects or a large number of individual defects separately. How best to group defects together for collective treatment can generally be determined from the nature of the inspection operations, whether visual or gauging, and the expectancy of defects as determined from the quality history Items involving visual inspection, can often be treated collectively to advantage.

As is generally true, the layout of an inspection plan depends to a considerable extent on the nature of inspection operations to be performed. Simplicity of administration is always to be desired. From the standpoint of minimizing overall inspection costs, it is often preferable, where several quality characteristics are to be inspected, to break down the inspection work into two or more separate inspection steps, each covering a relatively small number of characteristics.

III. Inspection of a Flow of Individual Lots or Sub-lots

15. Purposes of Inspection. A manufacturer's inspection of his own product serves two purposes[9].

(a) *Process Control*—To provide a basis for action with regard to the production process with a view to better future product.

(b) *Product Acceptance*—To provide a basis for action with regard to the product already at hand

The plan outlined in Part II has both of these purposes in mind, but the provision for selecting sample units continuously from the production line places special emphasis on control It aids, for example, in the prompt detection of defects and location of causes of trouble in the manufacturing process.

[9] See A. S. A War Standard, Z1. 3, *Control Chart Method of Controlling Quality During Production*, pp 5–6, 1942, American Standards Association, New York

The problem of acceptance of product is often eased, though at some sacrifice to the control aspects of the inspection work, if product is submitted to the inspector in lots or sub-lots and a sample taken from each.

16. Inspection procedure for sub-lots. With minor modifications, the plan and procedure of Part II can be extended to the case where material is offered as a *flow of consecutive sub-lots* of articles. In the inspection of parts, for example, the material may be offered in pan-loads or trays, each containing a collection of parts produced under essentially the same conditions. Or again, the product from a common source for a given short period of time, such as a half-hour, one hour, etc., may often be treated as a sub-lot and offered to the inspector as such for his acceptance. In what follows, however, it is essential that such sub-lots be kept in the *order of their production.*

The theoretical development given in Part II makes use of random-order spacing of defects in a statistically controlled product, with the specific provision that the units inspected be selected in the order of their production. In applying the general plan to the inspection of a *flow of consecutive sub-lots*, we no longer have individual units available in the order of their production. But we can use the same theoretical framework if we consider the random spacing of defects as their spacing in the chain of inspected units arranged in the *order of their inspection.* The probability distribution of the spacing of defects in inspected units will be the same regardless of the manner of selecting the units to be inspected, so long as we hold to the concept of statistical control in our solution.

The "i units in succession to be found clear of defects," discussed in Part II will now be defined as i *consecutively inspected units.* During sampling inspection, a group sample of units will be selected from each sub-lot, and the fraction f will relate to the ratio of the number of units in the sample to the total number of units in the sub-lot The fraction f will be held constant for all sub-lots. Furthermore, when it is required under the general plan to find i inspected units in succession clear of defects, the 100% inspection must be allowed to extend into immediately succeeding sub-lots if i units in succession are not found clear in the current sub-lot.

17. Procedure B. The procedure is as follows:

(a) At the outset, start inspecting 100% of the units in a sub-lot and continue such inspection until i inspected units in succession are found clear of defects. Extend the 100% inspection, if necessary, into one or more succeeding sub-lots in the order of their production.

(b) When i inspected units in succession are found clear of defects, discontinue 100% inspection and inspect only a fraction f of the units from each of the sub-lots, selecting the sample units in such a way as to fairly represent the sub-lot.

(c) If a sample unit is found defective, start a 100% inspection of the remainder of the sub-lot, and continue the 100% inspection until again i

inspected units in succession are found clear of defects, as in paragraph (a), extending such inspection into succeeding sub-lots, if necessary.

(d) In the event the 100% inspection extends into one or more succeeding sub-lots, if the number of units inspected in the last of such succeeding sub-lots exceeds a fraction f of the number of units in the sub-lot, accept this last sub-lot without further inspection. If on the other hand, the number of units inspected in this last sub-lot is less than the fraction f, inspect additional units from this same sub-lot to make up a sample equal to a fraction f of the number of units in the sub-lot.

(e) Correct or replace with good units all defective units found.

As was the case in Part II, the inspection plan is defined by two constants, f and i, and the protection offered is expressed in terms of AOQL. This sub-lot inspection plan differs from those already published in that the screening action is not confined to a single sub-lot but may extend over a succession of sub-lots, the entire production being regarded as a train of sub-lots that are linked together for purposes of inspection in the order of their production.

IV Remarks

It will have been noted that the plan here outlined should be regarded as a "special purpose" plan applicable under the conditions which have been enumerated—where production is practically continuous, where inspection is to be made during production or immediately thereafter and is to serve not only as a screening *acceptance* agency if necessary, but as an aid to *process control* by disclosing promptly any sub-standard quality conditions in the product. It is believed that the general plan provides a structure, which with possible variations in procedure to serve particular circumstances, may be found useful in designing additional sampling inspection techniques.

ON THE THEORY OF RUNS WITH SOME APPLICATIONS TO QUALITY CONTROL[1]

By J. Wolfowitz

Columbia University

1. Recent developments in the theory of runs. The increasing number and importance of recent advances in the theory and statistical applications of runs may make a brief paper on the subject of some interest The large volume of material and its wide dispersal, together with the limitations of space, will of necessity make these remarks far from exhaustive and complete

I shall not define a run because new advances and applications of new criteria to new problems would probably soon render most definitions obsolete. Runs as used in statistics are best characterized by a philosophy and a technique rather than by the employment of any one specific device. What is always involved is the ordering of observations according to some characteristic and the resultant effect of this ordering on the ordering according to some other characteristic. For example, if the seats at a meeting of statisticians and engineers are numbered and occupied by m engineers and n statisticians, then if we list the numbers of the occupied seats in ascending order and replace each number by E or S according as the seat is occupied by an engineer or statistician, we shall have a sequence of $m + n$ elements, m E's and n S's Thus, if $m = 7$ and $n = 6$, such a sequence might be

$$E\,E\,E\,S\,E\,E\,S\,S\,S\,E\,S\,S\,E.$$

If we were interested in knowing how well engineers and statisticians are acquainted with one another, we should find it of interest to study the runs of E's and S's in this sequence. Any subsequence of consecutive E's or S's which cannot be enlarged is called a run. Thus in the example above there is a run of E's of length 3, followed in order by a run of S's of length 1, a run of E's of length 2, a run of S's of length 3, a run of E's of length 1, a run of S's of length 2, and a run of E's of length 1. Runs of this kind are usually called runs of two kinds of elements. Naturally the characteristic according to which we order (in the example above, seat number) and the characteristic whose runs are observed (E or S) may be various. They ought in general to have a meaningful connection.

The order of observations has no value if it is known that the observations are independent and random from the same universe and one seeks to estimate a parameter of the universe Many of the statistical problems treated in the literature are of this character. In quality control of manufactured articles one

[1] Revised from an expository address delivered at a joint meeting of the Institute of Mathematical Statistics and the American Society of Mechanical Engineers at New York, May 29, 1943, at the invitation of the program committee

of the fundamental problems is to decide whether the observations are "random," or in the language employed in this field, whether statistical control exists For this purpose indiscriminate pooling of data which suppresses the order characteristics of the observations represents a loss of valuable information.

The algebra of runs of two kinds of elements is fairly elementary and most of the distribution problems involved have been solved Suppose an urn contains m white balls and n black balls, thoroughly mixed, and $m + n$ drawings are made without replacement. There are $\frac{(m + n)!}{m!n!}$ different sequences of W's and B's possible, and each sequence has the same probability. Let us find in how many ways the m elements W can be arranged to give k runs By a trick due to Euler, this is the coefficient of x^m in the purely formal expansion of

$$(x + x^2 + \cdots + x^m)^k$$

which is the same as the coefficient of x^m in the formal expansion of

$$(x + x^2 + x^3 + \cdots)^k = \left(\frac{x}{1 - x}\right)^k$$

and is therefore $\binom{m - 1}{k - 1}$ (which is, of course, the combinatorial symbol for $\frac{(m - 1)!}{(m - k)!\,(k - 1)!}$).

It is easy to see that the number of sequences of W's and B's which have $2k$ runs of both kinds is

$$2\binom{m - 1}{k - 1}\binom{n - 1}{k - 1}$$

and hence that the probability that U, the number of runs of both kinds, be $2k$ is

$$2\binom{m - 1}{k - 1}\binom{n - 1}{k - 1}\binom{m + n}{m}^{-1}$$

The details of this and other relevant derivations can be found in Wilks [1], Mood [2], Wald and Wolfowitz [3], and Stevens [12]. The formulae given there are of the type given above; e.g , for the probability that $U = c$ Application to tests of significance usually requires formulae of the type which give the probability that $U \leq c$. This causes some difficulty in application and raises a need for suitable tables Useful tables have been given by Swed and Eisenhart [4] and by P. S. Olmstead in an article by Mosteller [5]. The latter table really deals with a special case of runs of two kinds of elements.

The devices described above were systematically utilized by Mood [2] to give a valuable collection of formulae. A representative result is that the joint distribution of the numbers of runs of length 1, 2, $\cdots$, p and all those of length greater than p is asymptotically normal, with means and covariance matrix given.

The results given by Mood are limited to a classification of runs into a finite number of classes. The author [6] has given a general result which permits weighting runs of all lengths.

Closely allied to runs of two or more kinds of elements are runs from a binomial or multinomial population If the observations are classified into k classes, designated by $1, 2, \cdots, k$ say, and each observation has a constant probability p_i of falling into the ith class ($i = 1, 2, \cdots, k$) then a sequence of l observations all of which belong to the same class and which is preceded and followed by observations which belong to another class (except, of course, when the sequence is at the beginning or at the end of the series) is called a run of length l. If a coin, whether unbiassed or not, is tossed repeatedly, the runs of heads and tails are runs from a binomial population (i.e., $k = 2$) and if the coin is unbiassed, $p_1 = p_2 = \frac{1}{2}$.

The algebra of these runs has been studied mainly by von Bortkiewicz [7], von Mises [8], Wishart and Hirshfeld [9], Cochran [10], and Mood [2]. Runs from a binomial population (say) differ from runs of two kinds of elements in that m and n (defined above) are chance variables. If therefore, in general, a distribution formula valid for a fixed m and n be multiplied by the probability of this particular set of m and n $\left(\binom{m+n}{m} p_1^m p_2^n\right)$ and summed over m and n, the result will be the corresponding distribution formula for runs from a binomial population. Von Bortkiewicz [7], Cochran [10] and Mood [2] derived the essential parameters involved. Wishart and Hirshfeld [9] proved the asymptotic normality of the total number of runs from a binomial population, and these results were generalized by Mood [2].

Von Mises [8] proved that if N be the number of observations from a binomial population, the distribution of the number of runs of a length which is of the order of log N approaches the Poisson distribution with increasing N.

Cochran [10], extending the work of Gold [11], made use of runs of this kind in order to study what they called "the persistence of weather", i.e., whether dry months tend to follow dry months and wet months to follow wet months. In a long series of weather observations the months were classified as wet or dry and a four-fold table constructed of the number of months falling into each of the following categories:

(a) wet month following a wet month
(b) wet month following a dry month
(c) dry month following a wet month
(d) dry month following a dry month.

The chi-square test was applied to the four-fold table to test the null hypothesis that the probability of whether a month was wet or dry was independent of what its predecessor had been.

Olmstead [13] has made use of a run which is very similar to that of a run from a binomial population, except that the sequence terminates whenever an observation on a specified one of the two classes (a "failure") is recorded. The author

[6] has used a run defined as a sequence of consecutive integers in a permutation of the first n integers to test whether two variates are independently distributed when nothing is known about their distribution functions except that they are continuous. The rank correlation coefficient is usually employed for this purpose.

Of great importance in quality control of manufactured output are runs up and down. If, in any of the $n!$ equally likely (by hypothesis) permutations of the first n integers, we subtract each element from its successor and replace the result by $+$ or $-$ according as the difference is positive or negative, we get runs of $+$ signs and $-$ signs, called respectively runs up and down. The usage of the term length varies; in this paper we shall say that the length of a run is the number of $+$ or $-$ signs in it. This has the advantage that then the sum of the lengths of all the runs is $n - 1$. (Most quality control literature, which follows Shewhart [14] and Kermack and McKendrick [15], defines the length of a run as one more than the number of $+$ or $-$ signs in it.) Thus, for example, the sequence

3 4 7 6 5 1 2

will appear as

$$+ + - - - +$$

after the $+$ and $-$ signs have been inserted, and has an ascending run of length 2, followed by a descending run of length 3, followed by an ascending run of length 1.

The distributions associated with runs up and down in general present mathematical difficulties greater than those associated with distributions of runs of two kinds of elements and the results are far from complete. The asymptotic expectation of r_p, the number of runs of length p, was given with great brevity by Fisher [16] and in detail by Kermack and McKendrick [15], and the exact result was supplied by Wallis and Moore [17]. The matrix of covariances among the runs of various lengths is being computed, and, it is hoped, will be available for publication shortly. As far as the author is aware, no explicit formulae giving the probability that $r_p = k$ or that $r_p < k$ are known. Some recursion formulae of limited usefulness are available.

The author has recently obtained the asymptotic distributions of r_p, of $r_{p_1}, r_{p_2}, \cdots, r_{p_k}$ jointly, and of related statistics. These are jointly normal. Hence certain quadratic forms in these variables have approximately the chi-square distribution.

Anticipating somewhat the discussion to be given below, it may be mentioned here that the quadratic forms in certain of the r_p which Kermack and McKendrick [15] use to test for randomness, do not have the chi-square distribution which Kermack and McKendrick imply to them. Wallis and Moore [17] first pointed out that these quadratic forms were not the proper chi-square statistics for goodness of fit because of correlation among the r_p. The author's recent results show that these forms do not have the chi-square distribution.

2. Remarks on applications of runs. Let us now turn to statistical applications of some of the runs described above. Suppose we have a sample of m random independent observations on one variate and a similar sample of n observations on another variate. Suppose further that nothing is known a priori about the distribution of each except that both are continuous, and it is desired to test whether the two distributions are identical. This problem is of great practical importance and occurs frequently. In quality control of manufactured output it may occur, for example, if we wish to test whether the output of two machines, two workers, two different processes, or that from raw material obtained from two different sources, is the same. Naturally the problem not only of two, but in general, of a larger number of samples may arise.

The solution proposed in [3] is as follows: Let the $m + n$ observations be arranged in order of, say, ascending size, and let each observation be replaced by F of S according as it comes from the first or second sample. The total number U of runs in both F and S is the statistic to be used. Small values of U are the critical values for rejecting the hypothesis of identity of distributions. Thus in the example above of the seating of statisticians and engineers in the auditorium, a small value of U, which implies that the S (statisticians) and the E (engineers) each tend to bunch together, would be regarded as evidence that the statisticians and engineers present are not well acquainted with one another.

The statistic U seems a not unreasonable one for the purpose. A discrepancy between the two distribution functions will make alternation of values of the two variates less frequent This idea was proved for large n in [3], where a generalized concept of statistical consistency is given

On the other hand, the choice of U as a statistic is arbitrary; other reasonable criteria can certainly be given (see, for example, Dixon [19]. In [3] it is shown that a criterion which had previously been proposed was not acceptable because the statistic was not consistent, but nevertheless consistency is a property enjoyed by many statistics and constitutes only a partial check on the arbitrariness of choice. An "abnormally" long run in one or both variates which would be regarded by "common sense" as an indication that the hypothesis ought to be rejected, might be accompanied by a large number of runs of length one which might make the value of U not critically low. Some writers suggest that the presence of a long run of sufficient length be regarded as indicating rejection of the null hypothesis. In that case, if most of the runs were comparatively long, while none were critically long, the null hypothesis would not be rejected under this criterion, but the value of U would be small A step has been made in the direction of setting-up a criterion for the choice of statistic ([6]) so as to remove this arbitrariness. This involves an extension of the likelihood ratio principle. It must be remembered, however, that almost any criterion will fail to reject some sequence which, it seems intuitively, ought to be rejected. All statistical inference involves risks of error, one object of the science of statistics is to minimize these risks.

Another possible test for the problem of two samples is to compare the num-

bers of runs of various lengths with their expected numbers by the proper chi-square (Caution: the correlation among the variates must be taken into account). The author [6] has developed another test from an extension of the likelihood ratio.

Whenever a uniformly most powerful test does not exist, and this is the case in most non-parametric problems, it is not usually possible to say that one test is more powerful than another, unless the set of alternatives is sufficiently delimited. The power function is then the ultimate criterion for the choice of statistic.

If a sequence of n unequal numbers be given, a very important question is to decide whether the sequence is a "random" one; if it is and the sequence represents measurements on a characteristic of successive products of some manufacturing process, the latter is said to be in statistical control. A precise mathematical formulation can be given to this statement about randomness. Let $X_1, X_2, \cdots, X_n$ be chance variables, and let $x_1, x_2, \cdots, x_n$ be a set of random observations on the corresponding variables. To test whether $x_1, x_2, \cdots, x_n$ is a "random" sequence means to test the hypothesis that $X_1, X_2, \cdots, X_n$ are independently distributed and have identical distribution functions. This is in general a difficult problem, chiefly because of the large class of alternatives to the null hypothesis

Since the null hypothesis does not specify the distribution functions but only asserts their identity, the tests most generally sought have been such that their size is independent of the unknown (but identical for all the chance variables) distribution function. Certain reasonable procedures have been based on the numbers and lengths of runs up and down in the sequence.

R. A. Fisher [16] suggested doing this, but gave no indication as to what statistic was to be used. Kermack and McKendrick [15] and Wallis and Moore [17] propose the following procedure, the former writers implicitly and the latter explicitly: Let

$$r'_p = \sum_{i=p}^{n-1} r_i$$

and denote by $\bar{x}$ the expectation of the general chance variable x. The proposed statistic is

$$\sum_{i=1}^{p-1} \frac{(r_i - \bar{r}_i)^2}{\bar{r}_i} + \frac{(r'_p - \bar{r}'_p)^2}{\bar{r}'_p}$$

with the critical region the upper tail. Wallis and Moore recommend $p = 3$ and approximate the distribution by empirical methods. As we have seen above, Kermack and McKendrick err in ascribing to the statistic the chi-square distribution.

The criticism has been made by Olmstead [19] that this statistic is insensitive to pronounced trends in the data. This is correct, and had been pointed out earlier in [17], where the prior removal of a trend is recommended. Since one of

the important problems of quality control is detection of a trend, this would limit the usefulness of the statistic for quality control purposes.

It happens frequently when a new rank statistic has been proposed for testing a non-parametric hypothesis such as that of "randomness" above, that critics of the proposed criterion construct sequences which, they say, appealing to "ordinary common sense", any reasonable statistic ought to place in the region of rejection for almost any size of test. They then cheerfully point to the fact that the proposed statistic does not act in this reasonable fashion. A few remarks about this may not be amiss.

A test for, say, "randomness", which is to be made on the sequence of ranks, is really a numbering of the $n!$ permutations of, say, the first n integers, according to the order in which they ought to be taken into the critical region in order to make the latter of any prescribed size. This numbering could even be done by tabulating, for different n, the various sequences in their proper order. Aside from the obvious practical obstacles to such a tabulation, there would soon arise the difficulty that, after the "obvious" sequences are assigned their places the investigator would have difficulty in assigning to most of the remainder an ordering according to the degree in which they may be held to "contradict" the null hypothesis. Resort is therefore made to a statistic which can be given as an analytic expression in the ranks. Because of the inadequacies of the theory the formula is often chosen by analogy with a similar formula in classical statistics. Difficulties may arise because of this.

Let us examine for a moment this intuitive notion of reasonableness. Most people, and even most statisticians, would agree that the sequence of the first n integers in ascending order is an indication of non-randomness. The basis for this notion is an intuitive conception of an alternative to the null hypothesis for which this sequence is very probable. The fact is, however, that if we admit all alternatives to the hypothesis of randomness, for any sequence of ranks whatever there exist infinitely many alternatives which assign to this sequence a probability of one.

It seems to us that the difficulty can be met to a large extent by delimiting the class of distributions which constitute the alternatives to the null hypothesis, and by assigning to the admissible alternatives a weight function which measures the importance of the various alternatives (e.g., the financial loss caused by each). A profound treatment of this subject for the parametric case has been given by Wald [20]. This method has also the great merit that it removes the need for a choice of size of the region of rejection.

In the control of the quality of mass production output one of the outstanding problems is to decide on the basis of a sequence of observations on the product whether the production process is in statistical control. Shewhart and his school of industrial statisticians base many of their tests on the sequence of ranks. On the basis of their experience they find that the causes which most often lead to a breakdown of statistical control are such as to cause shifts up and down in the level of the observations or trends in the observations. To detect

the former they have devised the technique of runs above and below the median and to detect the latter they use runs up and down. Runs above and below the median may be described briefly as follows: The $2m + 1$ (odd number) of observations furnish a sequence of rankings from 1 to $2m + 1$. The elements 1 to m are considered to be elements of one kind and the elements $m + 2$ to $2m + 1$ elements of another kind. We then have a special case of runs of two kinds of elements. Limitations of space prevent the presentation of more detail or a description of the ingenious scheme by which both kinds of runs are graphically exhibited. The reader is referred to [14], [5], and [21], among others. The tests used in the industrial applications are not always explicitly stated, nor do they always seem to be the same. The most common involve comparison of runs of various lengths with their expected number or else are based on the presence of abnormally long runs.

A pretty application of the theory may be found in Campbell [21]. The corrosion of a copper plate was determined by a delicate mechanism which measured the electrical resistance in various places on the plate The rectangular plate was divided by rows and columns into forty small rectangles in each of which a measurement was made. The readings were made in each column in successive order from one end to the other, and the columns were also measured in successive order from one edge to the other. The observations, when examined for runs above and below the median and runs up and down, indicated something amiss ("absence of statistical control") Two causes were considered possible:

(a) variations, over the plate, in the corrosion of the copper;

(b) malfunctioning of the delicate measuring apparatus.

The runs obtained by arranging the observations in successive order according to positions on the plate might be expected to be associated with (a), while the runs obtained by arranging the observations in temporal order might be expected to be associated with (b). The object was therefore to separate the two orderings and this was done as follows: The rectangles were numbered 1 to 40 in the order in which the first observations had been made and a random permutation of this sequence was used to indicate the order in which the next set of observations was to be made. The second set was then ordered in two different ways, first according to the temporal order of the observations, and second according to the original ordering by positions. The runs above and below the median and the runs up and down, in the first ordering of the second set of observations gave evidence of a lack of statistical control, while those in the second ordering of the same set did not. An investigation located the trouble in the measuring apparatus.

3. Conclusion. The manifold achievements of quality control as it is practiced at present point to the desirability of still further development of theory and practice. We conclude this paper by suggesting a few directions in which the theory of runs could develop and be of greater assistance in quality control.

(1) The kinds of runs and the statistics used for making decisions in a production process should be chosen on the basis of the kind of deviations from the

state of statistical control which the engineers consider most likely to occur. It is very likely that different production processes may require different statistical procedures.

(2) General distribution theorems should be developed, power functions obtained, and the correlations between different tests investigated.

(3) The application of the weight function idea of minimizing financial losses should be considered.

In these developments both engineers and mathematical statisticians would have important and complementary roles. The tempo of progress will depend in large part on the cooperation between them.

REFERENCES

[1] S. S. Wilks, *Mathematical Statistics*, Princeton, 1943
[2] A. M. Mood, *Ann. Math. Stat.*, 11 (1940), p. 367.
[3] A. Wald and J. Wolfowitz, *Ann. Math. Stat.*, 11 (1940) p. 147.
[4] Frieda S. Swed and C. Eisenhart, *Ann. Math. Stat.*, 14 (1943) p. 66
[5] Frederick Mosteller, *Ann. Math. Stat.*, 12 (1941) p. 228
[6] J. Wolfowitz, *Ann. Math. Stat.*, 13 (1942) p. 247
[7] L. von Bortkiewicz, *Die Iterationen*, Berlin, 1917.
[8] Richard von Mises, *Zeitschrift fur angewandte Mathematik und Mechanik*, 1 (1921), p. 298.
[9] J. Wishart and H. O. Hirshfeld, *Journal of the London Math. Soc.*, 11 (1936), p. 227
[10] W. G. Cochran, *Quarterly Journal of the Roy. Meteorological Society*, 64 (1938) p. 631.
[11] E. Gold, *Quarterly Journal of the Roy. Meteorological Society*, 55 (1929), p. 307
[12] W. L. Stevens, *Annals of Eugenics*, 9 (1939), p. 10
[13] P. S. Olmstead, *Ann. Math. Stat.*, 11 (1940), p. 363.
[14] W. A. Shewhart, "Contribution of statistics to the science of engineering", *Proceedings of the Bicentennial Celebration of the Univ. of Penna.*, Philadelphia, 1941.
[15] W. O. Kermack and A. G. McKendrick, *Proc. Roy. Soc.*, Edinburgh, 57 (1937), pp. 228-240, 332-376
[16] R. A. Fisher, *Quarterly Journal of the Roy. Meteorological Soc.*, 52 (1926), p. 250
[17] W. A. Wallis and G. H. Moore, *A Significance Test for Time Series Analysis*, New York, 1941.
[18] P. S. Olmstead, *Journal of the American Stat. Assn.*, 1942, p. 152
[19] W. J. Dixon, *Ann. Math. Stat.*, 11 (1940), p. 199.
[20] A. Wald, *Ann. Math. Stat.*, 10 (1939), p. 299
[21] W. E. Campbell, "Use of statistical control in corrosion and contact resistance studies", *Bell Tel. System Tech. Publications*, 1942.

THE ACCURACY OF SAMPLING METHODS IN ECOLOGY

By Paul G. Hoel

University of California at Los Angeles

1. Introduction. For a number of years journals on ecology have contained articles on sampling techniques for estimating the distribution of common species of plants in various regions. Although much experimental work has been done on this problem and although the problem is essentially statistical in nature, no theoretical work of any consequence seems to have been attempted. This paper considers the question of the relative accuracy of common sampling methods from a theoretical point of view by means of geometrical probability and statistical distribution theory.

There are three common methods of sampling used by ecologists. They are designated by the names of coverage, abundance, and frequency. For each of these methods of sampling, there are two common choices of sampling unit, namely, the quadrat and the transect. By the coverage of a species in a region is meant the total area covered by the projection on the ground of the crowns of the plants of this species. By abundance is meant the total number of plants of this species in the region By frequency is meant the number of sampling units in the region in which at least one plant of the species occurs. A quadrat is a sampling unit in the form of a square, usually several yards on a side. A transect is a sampling unit in the form of a straight line, coverage in this case being the length of line covered by the projection of the crowns.

In this paper it will be assumed that plants possess circular crowns Further, it will be assumed that plant species distribute themselves at random in the region to be sampled. This is not necessarily the case, since there is often a tendency for plants of a given species to distribute themselves at random or otherwise in groups rather than as single plants. However, if sampling units are somewhat comparable in size, the relative accuracy of these methods of sampling based on a random distribution would be expected to hold fairly well for distributions somewhat removed from this ideal situation. Further, by the proper choice of sampling unit size, some non-random distributions behave very much as though they were random.

The accuracy of a sampling method may be measured by the variance of the estimate of the quantity which is of interest. Here interest will be centered on the total coverage of a given species in the region being sampled. Thus, two sampling methods will be said to be equally accurate for coverage if they produce equal variances for the estimate of total coverage.

The quadrat unit of sampling will be considered first for the three methods of sampling, after which the transect unit will be considered.

2. Quadrat coverage. Let the region to be sampled be a square B units on a side. Let there be n quadrats, each a square A units on a side, distributed at random in the region. Finally, let the total number of plants of the species in question in the region be N, with the distribution of the radius of their crowns given by a frequency function $f(r)$ whose explicit form will be specified later.

First, consider a single plant of radius r and a single quadrat. The problem is to determine the variance of a, the area of that part of the plant lying in the quadrat. Now the probability that this plant will be found in any particular part of the region is obtained by treating the plant as a circle of radius r which is thrown at random in the region and then applying geometrical probability to the position of the center of the circle. Thus, considering only those situations when the center of the circle lies in the region, the probability that the circle will cover an area of at least $a > \frac{1}{2}\pi r^2$ units of the quadrat is given by the ratio of the area of the subregion inside the quadrat whose boundary is the locus of centers of circles of radius r which have precisely a units of their area inside the quadrat, to the area of the region. Probabilities of this type may be treated as functions of a. The expressions below for such probabilities follow directly from Fig. 1, which displays one corner of the quadrat.

$$
\begin{aligned}
&P_1[a < \text{area} < \pi r^2] = 4S_1/B^2, && a \geq \tfrac{1}{2}\pi r^2 \\
&P_2[0 < \text{area} < a] = 4S_2/B^2, && a < \tfrac{1}{2}\pi r^2 \\
&P_3[a = \pi r^2] = (A - 2r)^2/B^2, \\
&P_4[a = 0] = P_4.
\end{aligned} \tag{1}
$$

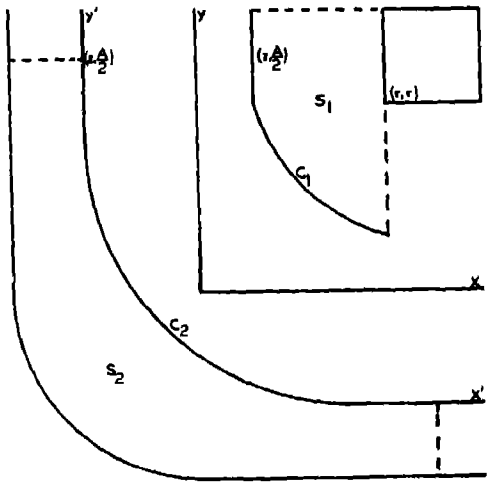


FIG. 1

Now

$$S_1 = (A - r)(r - z) - \int_z^r y\,dx, \tag{2}$$

where y is the ordinate of curve C_1. Likewise

$$S_2 = A(r + z) - z^2 + \tfrac{1}{4}\pi r^2 + \int_0^{-z+\sqrt{r^2-z^2}} y'\,dx', \tag{3}$$

where y' is the ordinate of curve C_2 with respect to the primed axes and z is negative. Using the formula for the area of a segment of a circle, the equation of C_1 is easily found to be

$$x\sqrt{r^2 - x^2} + r^2 \sin^{-1}\frac{x}{r} + y\sqrt{r^2 - y^2} + r^2 \sin^{-1}\frac{y}{r} = a, \qquad x^2 + y^2 \geq r^2 \tag{4}$$

$$x\sqrt{r^2 - x^2} + r^2 \sin^{-1}\frac{x}{r} + y\sqrt{r^2 - y^2} + r^2 \sin^{-1}\frac{y}{r} + 2xy + \tfrac{1}{2}\pi r^2 = 2a, \qquad x^2 + y^2 < r^2 \tag{5}$$

where the value of a is given in terms of z by

$$z\sqrt{r^2 - z^2} + r^2 \sin^{-1}\frac{z}{r} + \frac{\pi r^2}{2} = a. \tag{6}$$

The equation of C_2 is given by (5) with z negative These equations do not permit the solution of y in terms of x; however, they can be thrown into the following parametric form with t as parameter:

$$\begin{aligned} x &= r \sin\left\{\frac{t}{2} + \frac{1}{2}\cos^{-1}\left[\frac{a/r^2 - t}{\sin t}\right]\right\}, \\ y &= r \sin\left\{\frac{t}{2} - \frac{1}{2}\cos^{-1}\left[\frac{a/r^2 - t}{\sin t}\right]\right\}, \end{aligned} \tag{4'}$$

$$\begin{aligned} x &= r \sin\left\{\frac{t}{2} + \frac{1}{2}\cos^{-1}\left[\frac{2a/r^2 - \pi/2 + \cos t - t}{1 + \sin t}\right]\right\}, \\ y &= r \sin\left\{\frac{t}{2} - \frac{1}{2}\cos^{-1}\left[\frac{2a/r^2 - \pi/2 + \cos t - t}{1 + \sin t}\right]\right\}. \end{aligned} \tag{5'}$$

Since a may be treated as a parameter, equations (4) and (5), and hence (4') and (5'), represent a system of curves C_1 and C_2. Unfortunately, equations (4') and (5') are not convenient for integration purposes either, but they are convenient for numerical work. This system of curves can be approximated satisfactorily by means of simpler curves. One set of such approximating curves is the following system of circles:

$$(x - r)^2 + (y - r)^2 = (r - z)^2, \qquad z \geq 0 \tag{7}$$

$$(x - \sqrt{r^2 - z^2})^2 + (y - \sqrt{r^2 - z^2})^2 = (-z + \sqrt{r^2 - z^2})^2, \qquad z < 0. \tag{8}$$

Although inequalities may be obtained between the approximating and true curves, these are of little value for determining the accuracy of essential moments

obtained by using these approximating curves; therefore the accuracy of these approximating curves will be judged empirically by means of Fig. 2 in which the true curves are plotted by means of (4′) and (5′) for $z = .6, .3, 0, -.3, -.6, -.9$, of r with solid lines and the approximating circles (7) and (8) with broken lines. Although the circles appear to fit poorly for relatively large positive values of z, this is not serious because these values occur increasingly less often than other values of z for a random circle and because the use of these circles is confined to the rate of change of area bounded by these curves and the lines $x = r$ and $y = r$. Since the true curves are approaching the circles with decreasing positive z, their rate of change of area would not differ much from that for the circles even though the circles include larger areas for a given z. In the paragraph following (11), further evidence will be presented to show that for the computation of the first two moments of a, these curves give a good approximation.

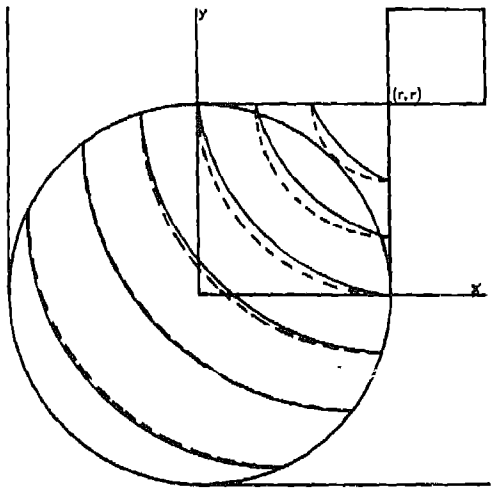


FIG. 2

For the purpose of obtaining the variance of a, consider the expected value of a^k. Since the variable a may be thought of as the sum of three variables which assume only the values 0, πr^2, and $0 < a < \pi r^2$, from (1) it follows that

$$E(a^k) = (\pi r^2)^k \frac{(A - 2r)^2}{B^2} + \int_{\frac{1}{2}\pi r^2}^{\pi r^2} a^k f_1(a)\, da + \int_0^{\frac{1}{2}\pi r^2} a^k f_2(a)\, da,$$

where $f_1(a)$ and $f_2(a)$ are the frequency functions for $z \geq 0$ and $z < 0$ respectively. Now since

$$P_1[a < \text{area} < \pi r^2] = \int_a^{\pi r^2} f_1(a)\, da,$$

and

$$P_2[0 < \text{area} < a] = \int_0^a f_2(a)\, da,$$

it follows from (1), (2), and (3) that

$$f_1(a)\,da = -dP_1 = -4\frac{dS_1}{B^2} = \frac{4}{B^2}\left[A - r + \frac{d}{dz}\int_z^r y\,dx\right]dz,$$

and

$$f_2(a)\,da = dP_2 = 4\frac{dS_2}{B^2} = \frac{4}{B^2}\left[A - 2z + \frac{d}{dz}\int_0^{-z+\sqrt{r^2-z^2}} y'\,dx'\right]dz.$$

Using the approximating curves (7) and (8), these integrals become:

$$\int_z^r y\,dx = r(r-z) - \frac{\pi}{4}(r-z)^2,$$

$$\int_0^{-z+\sqrt{r^2-z^2}} y'\,dx' = \left(1 - \frac{\pi}{4}\right)(-z + \sqrt{r^2-z^2})^2.$$

Hence,

$$f_1(a)\,da = \frac{4}{B^2}\left[A - 2r\left(1 - \frac{\pi}{4}\right) - \frac{\pi}{2}z\right]dz,$$

and

$$f_2(a)\,da = \frac{4}{B^2}\left[A - 2z - 2\left(1 - \frac{\pi}{4}\right)\left(2\sqrt{r^2-z^2} - \frac{r^2}{\sqrt{r^2-z^2}}\right)\right]dz$$

Hence,

$$E(a^k) = (\pi r^2)^k\frac{(A-2r)^2}{B^2} + \frac{4}{B^2}\int_0^r a^k\left[A - 2r\left(1 - \frac{\pi}{4}\right) - \frac{\pi}{2}z\right]dz$$
$$+ \frac{4}{B^2}\int_r^0 a^k\left[A - 2z - 2\left(1 - \frac{\pi}{4}\right)\left(2\sqrt{r^2-z^2} - \frac{r^2}{\sqrt{r^2-z^2}}\right)\right]dz.$$

Substituting the value of a from (6), standard integrals give the following values for $k = 1$ and $k = 2$:

$$E(a) = \frac{\pi r^4}{B^2}\left[\left(\frac{A}{r}\right)^2 + .13\right], \tag{9}$$

$$E(a^2) = \frac{\pi^2 r^6}{B^2}\left[\left(\frac{A}{r}\right)^2 - 1.15\left(\frac{A}{r}\right) + .46\right], \tag{10}$$

where certain constants involving π have been evaluated to two decimals.

If circles with centers outside the region but overlapping the region were also measured, then geometrical probability would give the following value for $E(a)$:

$$E(a) = \frac{\pi r^4}{B^2}\left(\frac{A}{r}\right)^2. \tag{11}$$

Since in (1) only circles with centers inside the region are assumed measured, $E(a)$ will be only very slightly larger than this last value; consequently the approximate result in (9) is only slightly in error. For a quadrat ten yards on a side and plants averaging two yards in diameter, the error is in the neighborhood of one tenth of one percent; consequently formula (10) may be expected to be quite accurate as well. Another approximating system of curves lying largely on the opposite side of the true curves from the circles gave formula (10) with .46 replaced by .26, both of which have a negligible effect on $E(a^2)$ for ordinary applications.

Formula (10) was derived on the assumption that the same circle was thrown repeatedly at random in the region. Consider now the situation when the circle varies in size according to the frequency function $f(r)$. Treating a and r as two statistical variables, their joint frequency function may be expressed as:

$$f(a, r) = f(r)f(a \mid r),$$

where $f(a \mid r)$ is the frequency function of a when r has the fixed value r. Letting $\mathfrak{E}(a^k)$ represent the expected value of a^k when r is permitted to vary according to $f(r)$,

$$\begin{aligned}\mathfrak{E}(a^k) &= \int\int a^k f(a, r)\, da\, dr \\ &= \int f(r) \int a^k f(a \mid r)\, da\, dr \\ &= \int f(r) E(a^k)\, dr,\end{aligned}$$

where all integrals are taken over the regions for which a and r are defined. Consequently, from (10) and (11)

$$\mathfrak{E}(a^2) = \frac{\pi^2}{B^2}\left[A^2\nu_4 - 1.15A\nu_5 + .46\nu_6\right],$$

and

$$\mathfrak{E}(a) = \frac{\pi}{B^2} A^2\nu_2 , \tag{12}$$

where the ν's represent moments of r. Hence the variance of a is given by:

$$\sigma_a^2 = \frac{\pi^2}{B^2}[A^2\nu_4 - 1.15A\nu_5 + .46\nu_6 - A^4\nu_2^2/B^2]. \tag{13}$$

Finally, let there be n quadrats, N circles whose radii vary according to $f(r)$, and let the total area of quadrat covered by the N circles be denoted by s. Then

$$\mathfrak{E}(s) = nN\mathfrak{E}(a), \tag{14}$$

and

$$\sigma_s^2 = nN\sigma_a^2 \tag{15}$$

The purpose of measuring s is to use it to obtain an estimate of T, the total area of the N circles. But

$$T = N\mathfrak{E}(\pi r^2) = N\pi\nu_2 . \tag{16}$$

Substituting the value of ν_2 from (12) and using (14),

$$T = B^2\mathfrak{E}(s)/nA^2.$$

Hence an estimate of T will be given by

$$T_1 = B^2 s/nA^2. \tag{17}$$

Using (15) and (13), the variance of this estimate will be given by

$$\sigma_{T_1}^2 = \frac{\pi^2 B^2 N}{nA^2}\left[\nu_4 - 1.14\frac{\nu_6}{A} + .46\,\frac{\nu_6}{A^2} - \frac{A^2}{B^2}\,\nu_2^2\right]. \tag{18}$$

3. Quadrat abundance. In this method the sampler merely counts the number of plants of the given species in each quadrat. Although this method was designed to estimate the total number of plants, it may be adapted to estimate total coverage as well. Since it is the practice to count a plant as lying in the quadrat only if its stem is in the quadrat, the probability that this event will occur is given by:

$$P_q = A^2/B^2. \tag{19}$$

Since there are n quadrats and N circles, the number of circles with centers lying in quadrats, which will be denoted by s, will follow the binomial distribution; hence

$$\mathfrak{E}(s) = nNP_q , \tag{20}$$

and

$$\sigma_s^2 = nNP_q(1 - P_q). \tag{21}$$

From (16) and (20) it follows that

$$T = \pi\nu_2\mathfrak{E}(s)/nP_q .$$

Therefore an estimate of T will be given by

$$T_2 = \pi B^2 m_2 s/nA^2, \tag{22}$$

where m_2 is a sample estimate of ν_2 obtained by measuring the diameters of k plants and calculating their mean area. Since m_2 and s are independent, a standard formula for the variance of a product of two independent variables may be applied to give

$$\sigma_{T_2}^2 = \left[\frac{\pi B^2}{nA^2}\right]^2 [\mathfrak{E}(m_2^2)\sigma_s^2 + \mathfrak{E}^2(s)\sigma_{m_2}^2].$$

But

$$\sigma^2_{m_2} = \frac{\nu_4 - \nu_2^2}{k},$$

and

$$\mathfrak{E}(m_2^2) = \frac{\nu_4 - \nu_2^2}{k} + \nu_2^2.$$

Consequently, with the aid of (19), (20), and (21)

$$\sigma^2_{T_2} = \pi^2 \left\{ \frac{N}{n} \frac{B^2 - A^2}{A^2} \left[\nu_2^2 + \frac{\nu_4 - \nu_2^2}{k} \right] + \frac{N^2}{k} [\nu_4 - \nu_2^2] \right\}. \tag{23}$$

4. Quadrat frequency. In this method the sampler records the number of quadrats observed and the number of those quadrats which contained at least one plant of the given species. Given N plants, the probability p that at least one of them will be found in a given quadrat is given by

$$p = 1 - (1 - P_q)^N,$$

where P_q is given in (19). For n quadrats the expected number of quadrats in which at least one plant will be found is therefore np. Letting w represent the number of such quadrats,

$$\mathfrak{E}(w) = n[1 - (1 - P_q)^N].$$

Solving for N,

$$N = \log\left[1 - \frac{\mathfrak{E}(w)}{n}\right] \Big/ \log [1 - P_q].$$

Consequently, from (16) an estimate of T will be given by

$$T_3 = \pi m_2 \log\left[1 - \frac{w}{n}\right] \Big/ \log [1 - P_q].$$

Neither the mean nor the variance of T_3 will exist because T_3 is a discrete variable which becomes infinite for $w = n$. Unless the density of the species is very low, values of w near n will occur quite often and hence cause T_3 to vary widely. Consequently the frequency method will be inferior to the abundance method except when the mean density is low, in which case the abundance method is practically as easy to apply. Because the frequency method is obviously inferior to the abundance method, it will not be considered further here.

5. Transect coverage. In this type of sampling a line is laid down and the length of line covered by a plant of the species in question is recorded. Let there be n such lines, each L units in length.

If a circle of radius r is thrown at random in the region, it will cross a line only if its center lies within the subregion, indicated in Fig. 3, composed of a

rectangle of width $2r$ and length L with semi-circular ends. From this figure it is clear that the probability of the circle intersecting some positive length less than z of the line is given by four times the shaded area s_3, divided by the area of the region. From this same diagram the following equations of the indicated curves result:

$$C_1: \quad \left(x - \frac{L}{2}\right)^2 + y^2 = r^2, \qquad \frac{L}{2} < x < \frac{L}{2} + r$$

$$C_2: \quad y = \sqrt{r^2 - z^2/4}, \qquad 0 \leq x \leq \frac{L}{2} - \frac{z}{2}$$

$$C_3: \quad \left(x - \frac{L}{2} + z\right)^2 + y^2 = r^2, \qquad \frac{L}{2} - \frac{z}{2} < x < \frac{L}{2} - z + r$$

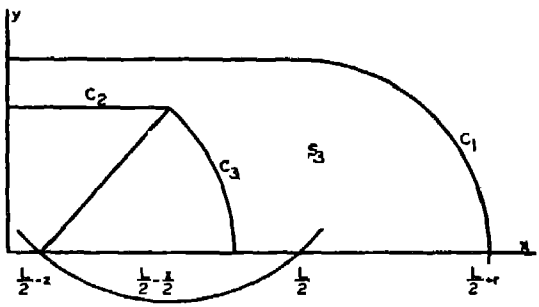


FIG. 3

Applying geometrical probability,

$$P_5[0 < \text{intercepted length} < z] = 4\frac{S_3}{B^2} = \int_0^z f(z)\,dz,$$

where $f(z)$ is the frequency function for z. But

$$S_3 = \frac{L}{2}[r - \sqrt{r^2 - z^2/4}] + \frac{z}{2}\sqrt{r^2 - z^2/4} + \frac{\pi r^2}{4} - \int_{\frac{1}{2}L-\frac{1}{2}z}^{\frac{1}{2}L-\frac{1}{2}z+r} \sqrt{r^2 - \left(x - \frac{L}{2} + z\right)^2}\,dx.$$

Standard integrals give

$$P_5 = \frac{4}{B^2}\left\{\frac{L}{2}[r - \sqrt{r^2 - z^2/4}] + \tfrac{3}{4}z\sqrt{r^2 - z^2/4} + \frac{r^2}{2}\sin^{-1}\frac{z}{2r}\right\}.$$

Consequently,

$$f(z)\,dz = dP_5 = \frac{4}{B^2}\left\{\frac{8r^2 + Lz - 3z^2}{8\sqrt{r^2 - z^2/4}}\right\}dz.$$

From this relation the following moments are readily obtained:

$$E(z) = \pi r^2 L/B^2$$

$$E(z^2) = [\tfrac{16}{3} L r^3 - \pi r^4]/B^2.$$

For variable r these formulas become:

$$\mathcal{E}(z) = \pi \nu_2 L/B^2,$$

$$\mathcal{E}(z^2) = [\tfrac{16}{3} L \nu_3 - \pi \nu_4]/B^2,$$

$$\sigma_z^2 = [\tfrac{16}{3} L \nu_3 - \pi \nu_4]/B^2 - \pi^2 \nu_2^2 L^2/B^4.$$

Let ξ denote total z for N circles and n quadrats, then

$$\mathcal{E}(\xi) = nN\pi \nu_2 L/B^2, \tag{24}$$

and

$$\sigma_\xi^2 = nN\{[\tfrac{16}{3} L \nu_3 - \pi \nu_4]/B^2 - \pi^2 \nu_2^2 L^2/B^4\}.$$

From (16) and (24)

$$T = B^2 \mathcal{E}(\xi)/nL.$$

Hence an estimate of T will be given by

$$T_4 = B^2 \xi/nL, \tag{25}$$

and its variance will be given by

$$\sigma_{T_4}^2 = \frac{N}{n}\left\{\frac{B^2}{L^2}[\tfrac{16}{3} L \nu_3 - \pi \nu_4] - \pi^2 \nu_2^2\right\}. \tag{26}$$

6. Transect abundance. Since the probability, P_t, of a circle of radius r intersecting a line of length L is the area of the band with semi-circular ends indicated in Fig. 3, divided by the area of the region,

$$P_t = [2rL + \pi r^2]/B^2.$$

Hence, letting s represent the total number of intersections, as in the case of quadrat abundance,

$$E(s) = nNP_t,$$

$$E(s^2) = nNP_t(1 - P_t) + n^2 N^2 P_t^2,$$

$$\mathcal{E}(s) = nN[2L\nu_1 + \pi \nu_2]/B^2, \tag{27}$$

$$\mathcal{E}(s^2) = \frac{nN}{B^4}\{B^2[2L\nu_1 + \pi \nu_2] + [nN - 1][4L^2 \nu_2 + 4\pi L \nu_3 + \pi^2 \nu_4]\}.$$

For simplicity of formula if $nN - 1$ is replaced by nN, the variance of s becomes

$$\sigma_s^2 = \frac{nN}{B^4}\{B^2[2L\nu_1 + \pi \nu_2] + nN[4L^2(\nu_2 - \nu_1^2) + 4\pi L(\nu_3 - \nu_1 \nu_2) + \pi^2(\nu_4 - \nu_2^2)]\}. \tag{28}$$

From (16) and (27)

$$T = \frac{\pi \nu_2 B^2 \mathfrak{E}(s)}{n[2L\nu_1 + \pi\nu_2]}.$$

Hence an estimate of T will be given by

$$T_s = \frac{\pi B^2 s}{n[\pi + 2L\alpha]},$$

where α is an estimate of ν_1/ν_2. In order to obtain a satisfactory estimate of ν_1/ν_2, data for the distribution of diameters of common California shrubs were analyzed. It was found that Pearson's type three curve gave an excellent fit. Since the moments of this type distribution are given by

$$\nu_m = \rho^m \prod_{j=1}^{m-1} [1 + jV^2], \tag{29}$$

where ρ is the mean and V is the coefficient of variation, σ/ρ, then $\nu_1/\nu_2 = 1/\rho\theta$, where $\theta = 1 + V^2$, and the above estimate becomes

$$T_s = \frac{B^2 s}{n}\left[1 - \frac{2L}{\pi\rho\theta + 2L} \cdot \frac{1}{1 + \varphi}\right], \tag{30}$$

where $\varphi = \pi\theta[\bar{r} - \rho]/[\pi\rho\theta + 2L]$ and where $1/\bar{r}$ is chosen as an estimate of $1/\rho$. Since $\bar{r}$ will be approximately normally distributed for samples considerably smaller than those usually taken to find $\bar{r}$, assume that it is normally distributed with mean zero and variance $\sigma^2\pi^2\theta^2/k[\pi\rho\theta + 2L]^2$. Since L is large relative to σ and since k will usually exceed twenty-five, this variance is very small, and hence the probability of φ exceeding one numerically is extremely small. Although the value $\varphi = -1$ is theoretically possible on the normality assumption, such a value would not permit the existence of either the mean or variance of $1/[1 + \varphi]$. However, if φ is restricted to a range of, say, ten standard deviations about zero, then $|\varphi| < 1$ for ordinary conditions and the variance will exist. Further, because φ assumes such small values, with this finite range the variance of $1/[1 + \varphi]$ is the same as the variance of φ itself if higher powers in this variance are neglected. Since s and φ are independent, the same product formula that was used for quadrat abundance may be employed here, together with the various approximations indicated above, to yield

$$\sigma^2_{T_s} = \left(\frac{N\pi\rho\theta}{2L + \pi\rho\theta}\right)^2 \left\{\left[\frac{B^2}{nN}(2L\nu_1 + \pi\nu_2) + 4L^2(\nu_2 - \nu_1^2) + 4\pi L(\nu_3 - \nu_1\nu_2) + \pi^2(\nu_4 - \nu_2^2)\right] \cdot \left[1 + \frac{4L^2V^2}{k(2L + \pi\rho\theta)^2}\right] + \frac{4L^2V^2}{k(2L + \pi\rho\theta)^2}[2L\nu_1 + \pi\nu_2]^2\right\}. \tag{31}$$

7. Comparison of methods. Formulas (18) and (23) may be compared for relative accuracy of these two quadrat methods of measuring coverage. Formulas (26) and (31) may be compared for relative accuracy of these two transect methods of measuring coverage. Finally, formulas (18) and (26), and formulas (23) and (31), may be compared to determine what length transect will give the same accuracy as a quadrat of given size All such comparisons will necessarily have to be done numerically by considering typical values for the parameters involved. The moments occurring in these formulas are expressible by means of (29) in terms of ρ and V if the form of $f(r)$ is that assumed here. For the data analyzed to determine $f(r)$ it was found that V was approximately 1/3. These numerical comparisons will not be made here.

The question of which type of sampling method should be employed now becomes a question of balancing relative ease or cost of sampling against size samples needed to produce equivalent accuracy as determined by means of these formulas. If total frequency is desired rather than total coverage, these formulas may be altered to handle this situation as well.

NEWS AND NOTICES

Readers are invited to submit to the Secretary of the Institute news items of general interest

Personal Items

Dr. Charles C Wagner has been named Assistant Dean of the School of Liberal Arts at the Pennsylvania State College

Miss Ruth E. Jolliffe has taken a position in the Graphic Analysis Department of Bell Aircraft Corporation.

Mr. H. F. Hebley has been appointed Director of Research for the Pittsburgh Coal Company

Lt F. W Dresch, USNR, U. S. Naval Proving Ground, Dahlgren, Virginia, has been promoted to the rank of Lieutenant Commander.

Mr. George F Mayer is a Sergeant in the United States Army and is stationed at Fort Lewis, Washington

Captain A. C Cohen, Jr. of Picatinny Arsenal has been promoted to the rank of Major.

New Members

The following persons have been elected to membership in the Institute:

Bassford, Horace R. B A (Trinity Coll.) Actuary, Metropolitan Life Insurance Co , 1 Madison Ave., New York, N Y

Benson, Kathryn E. M S (Washington) Teaching Asst., Univ of Calif , Berkeley, Calif.

Blackadar, Walter L. B A (McMaster) Asso Actuary, Equitable Life Assurance Society, 393 Seventh Ave., New York, N Y.

Buros, Asso. Prof. Oscar K. M A (Columbia) Rutgers Univ (on leave), Captain, Signal Corps, A U S *301 S. Courthouse Rd , Arlington, Va*

Clinedinst, William O. M.E (Carnegie Inst Tech) Eng , National Tube Co , Frick Bldg , Pittsburgh, Pa

Curry, Prof. Haskell B. Ph D. (Gottingen) Penna. State Coll., State College, Pa , *6708 N Sixth St., Philadelphia, Pa*

Dix, Margaret J. M.A (Rice Institute) Sec., Univ of Calif , Berkeley, Calif

Groth, Alton O. M.S (Iowa) Asst Actuary, Equitable Life Insurance Co of Iowa, Des Moines, Iowa.

Gurland, John M A (Toronto) Instr , Univ of Toronto, Toronto, Canada *97 Metcalfe St., Ottawa*

Humm, Doncaster G. Ph.D (Southern California) 1203 Commercial Exchange Bldg , 416 W. Eighth St., Los Angeles, Calif

Jahn, Fred S. M.S (Florida) General Manager, New Plastic Corp , 1017 N Sycamore, Hollywood, Calif

Jeming, Joseph M A. (Columbia) Captain, Army Air Corps *3010 Valentine Ave., New York, N. Y.*

Kavanagh, Arthur J. Ph.B. (Yale) Physicist, Spencer Lens Co , Buffalo, N Y. *19 Doat St.*

Kennedy, Evelyn M. M A. (Cincinnati) Industrial Economist, War Production Board, Washington, D C. *1462 Fairmont St , NW*

Lehmann, Eric L. M A (California) Asso., Univ. of Calif., Berkeley, Calif. *2514 Piedmont Ave.*

Lew, Edward A. M.A. (Columbia) Asst. Actuary, Metropolitan Life Insurance Co., 1 Madison Ave , New York, N Y.

Murphy, Ray D. A.B (Harvard) Vice Pres. and Actuary, Equitable Life Assurance Society, New York, N. Y *28 Godfrey Rd., Upper Montclair, N. J.*

Myers, James E. A B (Michigan) Leader-Statistical Analysis Group, Naval Res. Lab., Washington, D C *3014 Nichols Ave., SE.*

O'Connor, Harry W. M.B.A. (Harvard) Stat , Sperry Gyroscope Co Inc , Brooklyn, N. Y *37 Meadow Woods Rd , Great Neck.*

Painter, Frank M., Jr. M.B.A. (Harvard) Statistics Supervisor, Sperry Gyroscope Co , Brooklyn, N. Y. *343 82nd St*

Salkind, William M.B A. (Chicago) Asso Stat , U. S. Dept. of Agric., Washington, D C. *2149 K St , NW.*

Simon, Leon G. Pension Consultant *225 W. 34 St., New York, N Y*

Stewart, Oscar F. Statistics Supervisor, Sperry Gyroscope Co , Brooklyn, N. Y

Tucker, Ledyard R. B.S (Colorado) Res Asso , Univ. of Chicago, Chicago, Ill *5456 Greenwood Ave.*

Ullman, Joseph L. B A. (Buffalo) Teaching Fellow, Mass Inst. of Tech., Cambridge, Mass. *397 Jefferson Ave., Buffalo, N Y*

REPORT ON THE WASHINGTON MEETING OF THE INSTITUTE

The fifteenth meeting of the Institute of Mathematical Statistics was held at George Washington University, June 17–19, 1943 About 200 persons including the following sixty-one members of the Institute attended one or more of the three evening sessions:

T. W. Anderson, Jorge Arias, R O Been, H R. Bellinson, B M. Bennett, Richard Berger, Joseph Berkson, Felix Bernstein, Archie Blake, Dorothy S. Brady, W. G Cochran, J. B. Coleman, Gertrude Cox, J. H Curtiss, G B. Dantzig, Besse B Day, Robert Dorfman, H. F. Dorn, W. F. Elkin, W. D Evans, R. H Fadner, L. R Frankel, M A Girshick, Harry H. Goode, C. H Graves, T. N. Greville, F E. Grubbs, Louis Guttman, Morris H Hansen, W. A. Hendricks, W. N. Hurwitz, Walter Jacobs, Rachel M. Jenss, A. J King, G B. King, Lila F. Knudsen, H S. Konijn, Solomon Kullback, H G Landau, J E Lieberman, W G. Madow, Sophie Marcuse, J. W Mauchly, A. M. Mood, Harold Nisselson, Monroe L. Norden, H. W Norton, A. C. Rosander, David Rosenblatt, P. J Rulon, Marion Sandomire, W A. Shelton, Harry Shulman, J. H Smith, G W Snedecor, F. F Stephan, Alice Sternberg, Benjamin Tepping, J W. Tukey, C. R M. Tuttle, F M. Weida.

The following program, arranged by Dr. W. G. Madow, was held:

THURSDAY, JUNE 17 AT 8:00 P M.

APPLICATIONS OF SAMPLING THEORY

Chairman, PROFESSOR FRANK M WEIDA, George Washington University

1. *Some Recent Developments in the Application of Sampling Theory in Agriculture*
   Arnold J King, Iowa State College and Department of Agriculture; Walter A. Hendricks, North Carolina State College and Department of Agriculture
2. *The Relative Efficiency of Block Samples in Housing Surveys*
   Lester R. Frankel and William J. Cobb, Bureau of the Census
3. *The Optimum Size of Sampling Units*
   Dorothy Cruden and Alice Sternberg, Bureau of the Census

FRIDAY, JUNE 18 AT 8:00 P.M.

RECENT DEVELOPMENTS IN STATISTICAL THEORY

Chairman, PROFESSOR GEORGE W. SNEDECOR, Iowa State College

1 *On Some Recent Developments in Sampling Theory*
Morris H. Hansen, William N. Hurwitz, and William G. Madow, Bureau of the Census and Office of Price Administration
2. *On the Variance of Estimates Arising from Stratified Samples*
Frederick F Stephan, War Manpower Commission
3 *Statistical Techniques for the Comparison of Different Scales of Measurement*
William G. Cochran, Iowa State College
4. *Adjustments for Differential Refusal Rates in Samples of Human Populations*
Jerome Cornfield, Bureau of Labor Statistics
5 *On the Verification of Weather Forecasts*
Horace W Norton, Weather Bureau

SATURDAY, JUNE 19 AT 8.00 P.M.

SOME PROBLEMS IN STATISTICS

Chairman, COLONEL LESLIE E SIMON, War Department

1. *The Application of Statistical Methods in Acceptance Inspection*
Harold Bellinson, War Department
2. *The Distribution of the Radial Standard Deviation*
Captain Frank E. Grubbs, War Department
3 *Some Results in Tests of Randomness*
M. A Girshick, Department of Agriculture
4 *Corrections for Groupings*
John H Smith, Bureau of Labor Statistics
5. *On Group Blood Testing*
Robert E. Dorfman, Office of Price Administration

EDWIN G. OLDS,
Secretary

REPORT ON THE FIRST MEETING OF THE PITTSBURGH CHAPTER OF THE INSTITUTE

The first meeting of the Pittsburgh Chapter of the Institute of Mathematical Statistics was held at Carnegie Institute of Technology on Saturday, June 19, 1943. Thirty-six persons attended the meeting, including the following ten members of the Institute:

Shirley Bernstein, M A Brumbaugh, Karl Fetters, H. J Hand, G. E. Niver, F. G. Norris, E G. Olds, R. F. Passano, E M. Schrock, R. W. Shephard.

Morning and afternoon sessions were devoted to a round-table discussion of present industrial uses of statistical methods Mr. R. F. Passano, Bethlehem Steel Co., led the discussion Mr. F. G. Norris, Wheeling Steel Corp., acted as chairman of the sessions.

The Pittsburgh Chapter was formed from the Society of Industrial Quality Statisticians, which has held meetings at Carnegie Institute of Technology since 1941, with the object of providing a symposium for those interested in industrial applications. The Constitution of the Pittsburgh Chapter was ratified at the meeting. The object of the Chapter is to foster the advancement of mathematical statistics and to promote its application to industrial problems

The following officers for the Chapter for 1943 were elected:

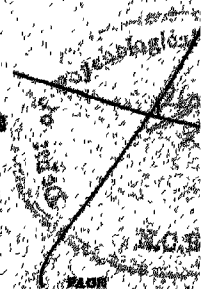
President, F. G. Norris, Wheeling Steel Corp.
Vice President, K. L. Fetters, Carnegie Institute of Technology
Sect.-Treas., H. J. Hand, National Tube Co.
Sponsor, E. G. Olds, Carnegie Institute of Technology
Board Members, R. F. Passano, Bethlehem Steel Co.
J. Manuele, Westinghouse Electric & Mfg. Co.

Howard Hand,
Secretary of the Pittsburgh Chapter

THE ANNALS *of* MATHEMATICAL STATISTICS

(FOUNDED BY H. C. CARVER)

THE OFFICIAL JOURNAL OF THE INSTITUTE OF MATHEMATICAL STATISTICS



Contents

PAGE

Statistical Inference in the Non-Parametric Case. HENRY SCHEFFÉ 305

On the Theory of Sampling from Finite Populations. MORRIS H. HANSEN AND WILLIAM N. HURWITZ 333

Multiple Sampling with Constant Probability. WALTER BARTKY 363

An Exact Test for Randomness in the Non-Parametric Case Based on Serial Correlation. A. WALD AND J. WOLFOWITZ 378

On a General Class of "Contagious" Distributions. W. FELLER 389

On the Construction of Sets of Orthogonal Latin Squares. H. B. MANN 401

On the Dependence of Sampling Inspection Plans upon Population Distributions. ALEXANDER M. MOOD 415

On Card Matching. T. W. ANDERSON 426

Notes:

The Detection of Defective Members of Large Populations. ROBERT DORFMAN 436

Further Points on Matrix Calculation and Simultaneous Equations. HAROLD HOTELLING 440

News and Notices 442

Report on the New Brunswick Meeting of the Institute 444

Report on the Second Meeting of the Pittsburgh Chapter of the Institute 445

Abstracts of Papers 445

Directory of the Institute 448

Vol. XIV, No. 4 — December, 1943

THE ANNALS OF MATHEMATICAL STATISTICS

EDITED BY

S. S. WILKS, *Editor*

A. T. CRAIG
W. E. DEMING
T. C. FRY
H. HOTELLING
J. NEYMAN
W. A. SHEWHART

WITH THE COÖPERATION OF

W. G. COCHRAN
J. H. CURTISS
H. F. DODGE
P. S. DWYER
C. EISENHART
W. K. FELLER
F. G. HOEL
W. G. MADOW
A. WALD

The ANNALS OF MATHEMATICAL STATISTICS is published quarterly by the Institute of Mathematical Statistics, Mt. Royal & Guilford Aves., Baltimore, Md. Subscriptions, renewals, orders for back numbers and other business communications should be sent to the ANNALS OF MATHEMATICAL STATISTICS, Mt. Royal & Guilford Aves., Baltimore, Md., or to the Secretary of the Institute of Mathematical Statistics, E. G. Olds, Carnegie Institute of Technology, Pittsburgh, Pa. Changes in mailing address which are to become effective for a given issue should be reported to the Secretary on or before the 15th of the month preceding the month of that issue. The months of issue are March, June, September and December. Because of war-time difficulties of publication, issues may often be from two to four weeks late in appearing. *Subscribers are therefore requested to wait at least 30 days after month of issue before making inquiries concerning non-delivery.*

Manuscripts for publication in the ANNALS OF MATHEMATICAL STATISTICS should be sent to S. S. Wilks, Fine Hall, Princeton, New Jersey. Manuscripts should be typewritten double-spaced with wide margins, and the original copy should be submitted. Footnotes should be reduced to a minimum and whenever possible replaced by a bibliography at the end of the paper; formulae in footnotes should be avoided. Figures, charts, and diagrams should be drawn on plain white paper or tracing cloth in black India ink twice the size they are to be printed. Authors are requested to keep in mind typographical difficulties of complicated mathematical formulae.

Authors will ordinarily receive only galley proofs. Fifty reprints without covers will be furnished free. Additional reprints and covers furnished at cost.

The subscription price for the ANNALS is $5.00 per year. Single copies $1.50. Back numbers are available at $5.00 per volume, or $1.50 per single issue.

COMPOSED AND PRINTED AT THE
WAVERLY PRESS, INC.
BALTIMORE, MD., U. S. A.

Entered as second-class matter at the Post Office at Baltimore, Maryland, under the Act of March 3, 1879

STATISTICAL INFERENCE IN THE NON-PARAMETRIC CASE[1]

By Henry Scheffé

Princeton University

CONTENTS

		Page
1.	Introduction	305
	Part I Non-parametric tests	
2	The randomization method of obtaining similar regions	307
3.	Goodness of fit. Randomness	312
4.	The problem of two samples	313
5.	Independence	316
6.	Analysis of variance	316
	Part II. Non-parametric estimation	
7.	Classical results on point estimation	320
8	Confidence intervals for an unknown median, for the difference of medians	321
9	Confidence limits for an unknown distribution function	322
10	Tolerance limits	323
	Part III. Toward a general theory	
11	The criterion of consistency	324
12	Likelihood ratio tests	325
13.	Wald's formulation of the general problem of statistical inference	327

1. Introduction. In most of the problems of statistical inference for which we possess solutions the distribution function is assumed to depend in a known way on certain parameters. The values of the parameters are unknown, and the problems are to make inferences about the unknown parameter values. We refer to this as the parametric case. Under it falls all the theory based on normality assumptions.

Only a very small fraction of the extensive literature of mathematical statistics is devoted to the non-parametric case, and most of this is of the last decade. We may expect this branch to be rapidly explored however: The prospects of a theory freed from specific assumptions about the form of the population distribution should excite both the theoretician and the practitioner, since such a theory might combine elegance of structure with wide applicability. The process of development will no doubt inspire some mathematical attacks of considerable abstractness. There are already signs that more number-theoretic problems and measure-theoretic problems will enter our subject through this door, and perhaps even some topological ones. Some ability to think in terms of

[1] Parts of this paper were used in an invited address given under the title "Statistical inference when the form of the distribution function is unknown" before the meeting of the Institute of Mathematical Statistics on September 12, 1943 in New Brunswick, N J.

functionals, function spaces, and metrization of function spaces will be useful in attempting general theories of "best" tests and estimates. Toward such abstract phases of the development the attitude of the practical statistician should be one of tolerance, for the new theory already promises to give him many new tools which are both simpler and of wider use.

While the maturity of the non-parametric theory is still in the future, it is well to remark that its beginnings go relatively far back. Of our most famous tests, such as Pearson's χ^2-test, Student's test, and Fisher's analysis of variance tests, the oldest concerns a non-parametric problem: In 1900 Karl Pearson proposed his χ^2-criterion to test the goodness of fit of a theoretical distribution to observations, and in 1911 he extended his χ^2-method to the problem of two samples. The first of these problems may be regarded as non-parametric if the choice of the theoretical distribution is not based on calculations from the data, and the second is without doubt a non-parametric problem. R. A. Fisher treated an analysis of variance problem non-parametrically at least as early as 1925, for in the first edition of his *Statistical Methods for Research Workers* we find the sign test. General formulations of the problems of statistical inference, and criteria for "good" and "best" solutions[2] have been advanced by R. A. Fisher, Neyman, E. S. Pearson, and Wald. These general theories were all strictly parametric until 1941 when Wald proposed one sufficiently broad to cover the non-parametric case.

We now introduce some notation to which we shall adhere throughout this paper. Statistical inferences are based on measurements. The total number of measurements will always be denoted by n. We conceive of n random variables $X_1, X_2, \cdots, X_n$ on which the measurements are made. The domain of each X_j can always be taken to be a set of real numbers. If vector random variables occur, the X_j will denote components. The cumulative distribution function (c.d.f.) of the random variables will be written $F_n(x_1, x_2, \cdots, x_n)$, — this is the probability that all $X_j \leq x_j$ simultaneously The c.d.f. F_n is then always defined in a complete n-dimensional Euclidean space W, called the sample space; W is the space of points $E = (x_1, x_2, \cdots, x_n)$. The sample point with the random coordinates $X_1, \cdots, X_n$ will be denoted by $\mathbf{E}$.

In describing the validity of specific non-parametric tests and estimates in the sequel it will be convenient to refer to the following classification[3] of *univariate* c.d.f's $F(x)$: Ω_0 is the class of all F. Ω_2 is the class of all continuous F. Ω_3 is the class of all absolutely continuous F, that is, those F for which there exists a probability density function $f(x)$, so that

$$F(x) = \int_{-\infty}^{x} f(\xi)\, d\xi.$$

Ω_4 consists of all F which may be written in the above form with f continuous.

[2] For a bibliography see [22].

[3] The notation follows [31].

PART I. NON-PARAMETRIC TESTS

2. The randomization method of obtaining similar regions. In any problem of statistical inference it is assumed that the c.d.f. F_n of the measurements is a member of a given class Ω of n-variate distribution functions, we write $F_n \in \Omega$. Ω is called the class of admissible F_n. If Ω is a k-parameter family of functions the problem is called parametric, otherwise, non-parametric. A statistical hypothesis H is a statement that $F_n \in \omega$, where ω is a given subclass of Ω. A test of the hypothesis H consists of choosing a Borel region w in the sample space W and rejecting H if and only if the sample point $\mathbf{E}$ falls in w; w is called the critical region of the test.

The choice of the critical region w is usually[4] made as follows: A positive constant α (ordinarily about .01 or .05) is chosen and called the significance level of test. If regions w exist for which $Pr\{\mathbf{E} \in w \mid F_n\}$—the probability that the sample point $\mathbf{E}$ fall in w, calculated from the c.d.f. F_n—is equal to α for all $F_n \in \omega$, then the choice of critical region is usually limited to this class. Such regions are very important in the theory of testing hypotheses, and it is convenient to have a name for them: Following the terminology of Neyman [22] in the parametric case we shall call them *similar* to the sample space with respect to all F_n in ω, or more briefly, similar regions. A similar region is then a region w for which $Pr\{\mathbf{E} \in w \mid F_n\}$ is the same constant for all $F_n \in \omega$. The advantage of using similar regions as critical regions is that the risk of rejecting the hypothesis when it is true (type I error) is controlled: no matter what member of ω the unknown F_n happens to be, the probability of rejection of the (true) hypothesis is exactly α. We remark here that the problem of the existence and structure of similar regions in the parametric case has been treated only under very heavy restrictions and must be considered still mostly unsolved, whereas we shall see later that in the non-parametric case it promises to be relatively simple.

When similar regions exist for a chosen α there is usually a large family of them. Ideally the choice of the critical region w from the family of similar regions would be based on a complete knowledge of two functionals of F_n for $F_n \in \Omega - \omega$, that is, for those F_n corresponding to the various admissible ways in which the hypothesis can be false. the first, the probability of rejection (of avoiding a type II error), namely $Pr\{\mathbf{E} \in w \mid F_n\}$, called the power function of w, and the second, the relative importance of rejection in the concrete situation in which the test is to be applied. In other words, one would like to choose the w with the power function "best" for the very specific problem at hand. However, little has been done along this line in the non-parametric case, and, as we shall note below, the choice of w from the family of similar regions is usually made by means of a statistic chosen on intuitive grounds.

A general method of obtaining similar regions, which we shall call the *randomization method*, will now be described. The credit for originating this method and envisioning its wide applicability belongs to R. A. Fisher, who first

[4] Another approach to the choice of critical region will be described in section 13.

used it in 1925 [3]. Consider the set S of permutations on the coordinates $x_1, x_2, \cdots, x_n$, which leave invariant all the c.d f's F_n in ω. Suppose the number of permutations in the set S is s; then s divides $n!$. Now define for any point E in W a corresponding set $\{E'\}$ of s points obtained by making on the coordinates of E the permutations of the set S. The value of the c.d.f. F_n is then the same at all s points E' generated by E, for all $E \in W$ and all $F_n \in \omega$ The s points of $\{E'\}$ will be distinct unless the point E lies in a certain region W_0 ; W_0 depends on the set S of permutations determined by the class ω, and will always be contained in the union of all diagonal hyper-planes $x_i = x_j$ $(i \neq j)$. A critical region w is constructed by the randomization method by choosing a positive integer $q < s$, and for every E not in W_0, putting q points of the corresponding set $\{E'\}$ in w and the remaining $s - q$ points outside w, by any rule whatever, just so w is a Borel set. We shall also say that a Borel set w is obtained by the randomization method if it has the structure just described except on a (Borel) subset w_0 of w having the property $Pr\{\mathbf{E} \in w_0 \mid F_n\} = 0$ for all $F_n \in \omega$ It may be shown by the methods used elsewhere [31] by the writer that if ω is a class of continuous c.d f's then the region w thus obtained is a similar region with $\alpha = q/s$; furthermore, that under mild restrictions (roughly, that the boundary of w be a sufficiently "thin" set), at least for certain classes ω, this is the *only* method of obtaining similar regions

One might call the set $\{E'\}$ of points corresponding to E the subpopulation of points "equally likely" under the null hypothesis H, but we shall call $\{E'\}$ simply the *subpopulation* corresponding to E The decision as to which q of the s points of the subpopulation are to be put into the critical region w is usually made with the aid of a statistic T chosen on an intuitive basis. By a statistic T we mean of course a function of the sample only, not depending on the c.d.f. F_n, thus $T(\mathbf{E}) = T(X_1, \cdot \quad, X_n)$. For a suitably chosen q, the q points of the subpopulation $\{E'\}$ giving $T(E')$ values in a certain range—usually the q largest or q smallest values—are put into w, and these q values are then called the "significant" values.

Before proceeding further let us consider an example illustrating all the definitions we have introduced thus-far. Suppose that on the basis of a sample of m pairs (X_i, Y_i), $i = 1, 2, \cdots, m$, from a bivariate population with unknown c.d f $G(x, y)$ we wish to test the independence of the random variables X, Y. To fit our general notation write $Y_i = X_{i+m}$. Assuming only that the sample is random, we have, with $n = 2m$, that the c d f. of the sample point $\mathbf{E}$ is of the form

$$F_n(x_1, \cdots, x_n) = \prod_{i=1}^{m} G(x_i, x_{i+m}).$$

Now suppose we know or are willing to assume further that the unknown c.d.f. $G(x, y)$ of the population is in a certain class $\Omega_\nu^{(2)}$ of bivariate c d.f.'s, where $\Omega_\nu^{(2)}$ is the bivariate analogue of the class Ω_ν of univariate c.d.f.'s defined in section

1; thus if we knew the unknown $G(x, y)$ were continuous, we would have $G \in \Omega_2^{(2)}$. The class Ω of admissible F_n is then

$$\Omega = \left\{ F_n \mid F_n = \prod_{i=1}^{m} G(x_i, x_{i+m}); G \in \Omega_\nu^{(2)} \right\},$$

where the notation $\{F_n \mid F_n \text{ of the form } \mathfrak{F}\}$ denotes the class of all F_n of the form $\mathfrak{F}$. The hypothesis of independence may now be expressed as $H: F_n \in \omega$, where the subclass ω of Ω is

$$\omega = \left\{ F_n \mid F_n = \prod_{i=1}^{m} F^{(1)}(x_i) \prod_{j=m+1}^{2m} F^{(2)}(x_j); F^{(k)} \in \Omega_\nu, k = 1, 2 \right\}$$

The set S of permutations which leave all $F_n \in \omega$ invariant is obtained by making all possible permutations of the first m coordinates $x_1, \cdots, x_m$ among themselves, and of the second m coordinates $x_{m+1}, \cdots, x_{2m}$ among themselves. The total number s of permutations in S is thus $(m!)^2$ Making these permutations on the coordinates of any point E in W, we get the set $\{E'\}$ of $(m!)^2$ points. The points of $\{E'\}$ are distinct unless E lies in the region W_0 defined as the union of all hyperplanes $x_i = x_j$ where $i \neq j$ and i, j are both in the set of integers $1, 2, \cdots, m$ or else both in the set $m + 1, \cdots, 2m$. Pitman [28] applied the randomization method to this problem, using as the statistic $T(E)$ the numerical value of the (sample) Pearsonian correlation coefficient,

$$T(E) = \left| \sum_{i=1}^{m} x_i x_{i+m} \right| \Big/ \left(\sum_{i=1}^{m} x_i^2 \sum_{j=m+1}^{2m} x_j^2 \right)^{\frac{1}{2}},$$

the large values of T being the significant ones We note that $T(E)$ takes on at most $m!$ different values over the subpopulation. What we previously called a "suitably chosen" q would be in the present case a multiple of $m!$, and the choice of significance level $\alpha = q/s$ would then be limited to multiples of $1/m!$

The method of randomization is seen to exploit whatever symmetry properties the F_n in ω possess as a class A special case of the general method is the *method of ranks* This gives regions of an especially simple form defined by certain inequalities on the coordinates. Probably the only case in which the method of ranks will ever be used is when the F_n in ω have the following special kind of symmetry: Suppose they are completely symmetrical in each of certain subsets of the coordinates, say t sets of $n_1, n_2, \cdots, n_t$ coordinates, respectively, where $\sum_{i=1}^{t} n_i = n$. We may assume the coordinates numbered so that F_n is completely symmetrical in the set $x_{p_i+1}, x_{p_i+2}, \cdots, x_{p_i+n_i}$ ($p_i = \sum_{j=1}^{i-1} n_j, i = 2, 3, \cdots, t; p_1 = 0$), for all $F_n \in \omega$. The set S of permutations is thus generated by making all $n_i!$ permutations on the n_i coordinates $x_{p_i+1}, \cdots, x_{p_i+n_i}$ ($i = 1, \cdots, t$), so that the total number of permutations in S is $s = n_1! \, n_2! \cdots n_t!$.

Corresponding to the i-th set of coordinates in which F_n is symmetrical, let us divide the sample space W up into $n_i!$ regions defined by

$$x_{p_i+1} < x_{p_i+2} < \cdots < x_{p_i+n_i}$$

and the $n_t! - 1$ other inequalities obtained by permuting the subscripts in the above. Denote these regions by $w_{t,k}$ $(k = 1, \cdots, n_t!)$. Let

$$w_{k_1,k_2,\cdots,k_t} = w_{1,k_1} \cap w_{2,k_2} \cap \cdots \cap w_{t,k_t},$$

that is, $w_{k_1,k_2,\cdots,k_t}$ is the part of W common to the regions $w_{1,k_1}, w_{2,k_2}, \cdots, w_{t,k_t}$. This process divides the sample space W up into s disjoint regions $w_{k_1,k_2,\cdots,k_t}$, which we shall now denote simply by w_σ $(\sigma = 1, \cdots, s)$. The set $\{w_\sigma\}$ of regions covers all of the sample space W except the region W_0 on which certain coordinates become equal. We shall say that the sample point $\mathbf{E}$ has the σ-th ranking, R_σ, if $\mathbf{E}$ falls in w_σ. We may then speak of a random variable $R = R(\mathbf{E})$, the "ranking", taking on the s possible values R_σ, or the "tied" ranking R_0 if $\mathbf{E} \in W_0$.

A critical region w is constructed by the method of ranks by taking w to be the union of q of the regions w_σ. Those rankings R_σ corresponding to the q regions w_σ constituting the critical region w, will be called the significant rankings. Any statistic $T(E)$ used as the criterion to decide which are the significant rankings now becomes a function of the ranking R only, $T(E) = U(R)$. We may regard the method of ranks as a simplification of the problem of testing statistical hypotheses, in which the infinite n-dimensional sample space W is replaced by a finite space of $s + 1$ points R_σ. If Ω is a class of continuous F_n we may ignore the point R_0 since then $Pr\{R = R_0\} = 0$.

In the problem of independence, which we have used before to illustrate the definitions of this section, the method of ranks was applied by Hotelling and Pabst [9], who took as the statistic $U(R)$ the numerical value of the Spearman coefficient of rank correlation, large values being significant.

The method of randomization yields similar regions if ω is a class of continuous functions. What will the method get us if we drop the continuity restriction? In this case we can no longer ignore the possibility that the sample point $\mathbf{E}$ fall in the exceptional region W_0, for we do not have $Pr\{\mathbf{E} \in W_0\} = 0$. We owe to Pitman [27] the following idea: We continue to use the subpopulation $\{E'\}$ and a chosen statistic $T(E)$ as above, but instead of separating the points of $\{E'\}$ into two classes (significant points and non-significant points) by means of $T(E)$ we now add a third class of "doubtful" points.[5] If the s points of the set $\{E'\}$ are not distinct they are to be counted according to their multiplicities under the process of applying the permutations of the set S to the coordinates of E. Suppose that the large values of T are significant. Number the s points of $\{E'\}$ so that $T(E'_1) \geq T(E'_2) \geq \cdots \geq T(E'_s)$. If $T(E'_q) > T(E'_{q+1})$ we call $E'_1, \cdots, E'_q$ significant, and the rest non-significant. However if $T(E'_q) = T(E'_{q+1})$, we term all points E' with $T(E') = T(E'_q)$ doubtful, points E' for which $T(E') > T(E'_q)$, significant, and points E' with $T(E') < T(E'_q)$, non-significant. This process divides the sample space W up into three regions instead of the customary

[5] Instead of the terms significant, non-significant, doubtful, Pitman uses discordant, concordant, neutral.

two, namely, a rejection region w_R, an acceptance region w_A, and a doubtful region w_D. It is a special case of the following procedure: For every set $\{E'\}$ define positive integers $m_R = m_R(\{E'\})$ and $m_A = m_A(\{E'\})$ such that $m_R \leq q$, $m_A \leq s - q$, and put m_R of the points E' in w_R, m_A of the points E' in w_A, and the remaining $s - m_A - m_R$ of the points E' in w_D, in any way so that w_R and w_A are Borel regions. When any E' is assigned to w_R or w_A it is to be counted according to its multiplicity as defined above, if $\{E'\}$ contains less than s distinct points. It may be shown that with $\alpha = q/s$, $Pr\{\mathbf{E} \epsilon w_R \mid F_n\} \leq \alpha$ and $Pr\{\mathbf{E} \epsilon w_A \mid F_n\} \leq 1 - \alpha$ for all $F_n \epsilon \omega$, that is, whenever H is true.

Before closing this section on the method of randomization, we mention a few difficulties which frequently arise when it is applied. Except for very small samples the calculation determining whether or not the observed value E_0 of the sample point $\mathbf{E}$ belongs to the significant points of the subpopulation $\{E_0'\}$ generated by E_0, is usually extremely tedious In such cases the author of the test often gives an approximation to the discrete distribution of his statistic $T(\mathbf{E})$ over the subpopulation $\{E'\}$ by means of some familiar continuous distribution for which tables are available, the laborious exact calculation by enumeration then being replaced by the computation of a few moments (that is, values of certain homogeneous polynomials in the observed coordinates) and the use of existing tables of percentage points of the continuous distribution [6] Barring some papers where the method of ranks is used, the justification of these approximations is never satisfactory from a mathematical point of view, the argument being based on a study of the behavior of two, or at most four, moments. The only exception to the last statement appears to be a very recent paper by Wald and Wolfowitz [42], which may point the way to genuine derivations of asymptotic distributions for the non-rank case of the randomization method. We shall distinguish between derivations of asymptotic distributions and arguments based on two or four moments by saying that a distribution is "proved" in the former case and "fitted" in the latter.

Another difficulty arises, most noticeably in the method of ranks, out of the possibility of equality of the observed coordinates In the distribution theory this is usually avoided by assuming ω to be a class of continuous c.d.f's, so that $Pr\{\mathbf{E} \epsilon W_0 \mid F_n\} = 0$ for all $F_n \epsilon \omega$, but in practice, since the measurements are usually made to about three significant figures, ties *do* occur in the sample. While some scattered work has been done on this question there is need for a thorough general treatment.

In some of the work that has been done on particular non-parametric tests

[6] In many cases the approximate test obtained by fitting a familiar distribution is found to coincide with widely used tests based on normality assumptions. In such cases if the fitting is asymptotically correct the following remarks are justified. (1) If the non-parametric test is used in a case where we hesitate to assume normality but normality actually exists, the non-parametric test is asymptotically as efficient as the older test assuming normality. (2) If normality is assumed when it does not exist, no error is incurred asymptotically when the older test is used.

it is not very clear just what the null hypothesis H is. Two situations often occur: Suppose $H: F_n \in \omega$ is the hypothesis we actually wish to test at significance level α. Let w be the chosen critical region and ω_w the class of F_n for which $Pr\{\mathbf{E} \in w \mid F_n\} = \alpha$. The two situations are (*i*) ω is a proper subset of ω_w, and (*ii*) ω_w is a proper subset of ω. Of these (*i*) seems less objectionable, for then the probability of a type I error (rejecting H when true) is strictly α, but the probability of accepting H is the same when certain alternatives ($F_n \in \omega_w - \omega$) are true as when H is true. In case (*ii*) the probability of a type I error is not α unless F_n is in the subclass ω_w of ω; thus there might be a much higher probability than α of rejecting H when it is true, if the true $F_n \in \omega - \omega_w$. To illustrate situation (*i*) consider K Pearson's χ^2-test for goodness of fit of a theoretical distribution $F_0(x)$ to a sample $\mathbf{E}$. Suppose $\mathbf{E}$ is from a univariate population whose true c.d.f. is $F(x)$. If F has the property that for the intervals I_j defined in section 3, $\int_{I_j} dF = \int_{I_j} dF_0$, $j = 1, 2, \cdots, N$, then the probability of rejection is the same as when the hypothesis is true. An example of (*ii*) might occur if we wish to test whether the *means* of two univariate populations are the same. If we use one of the tests of section 4 in which the probability of rejection is calculated under the assumption that the *distributions* of the populations are the same, then we do not know that the probability of a type I error is α, for the samples might come from two populations with the same mean but different distributions.

3. Goodness of fit. Randomness. The non-parametric case of testing goodness of fit is the following: On the basis of a sample $\mathbf{E}$ from a population with c.d.f. $F(x)$ known to be a member of some Ω_ν, we wish to test whether $F = F_0$, where F_0 is a given c.d.f. The class of admissible c.d.f.'s F_n is

$$\Omega = \left\{F_n \mid F_n = \prod_{i=1}^{n} F(x_i);\ F \in \Omega_\nu\right\},$$

and the hypothesis H specifies that $F_n \in \omega$, where

$$\omega = \left\{F_n \mid F_n = \prod_{i=1}^{n} F_0(x_i)\right\}.$$

K. Pearson's χ^2-test [25] consists of choosing an integer N, dividing the x-axis up into a set $\{I_j\}$ of disjoint intervals ($j = 1, 2, \cdots, N$), and using as statistic $T(E)$ the Pearsonian chi square,

$$\chi_P^2 = \sum_{j=1}^{N} [m_j - \mathfrak{E}(m_j)]^2/\mathfrak{E}(m_j),$$

where m_j is the number of observed coordinates of $\mathbf{E}$ in I_j, and $\mathfrak{E}(m_j) = n\int_{I_j} dF_0$. Large values of χ_P^2 are regarded as significant. Exact significance

levels for χ_P^2 could be obtained by considering its distribution over the subpopulation $\{E'\}$ generated by the sample This process would lead to the multinomial distribution of the m_j mentioned in the usual derivations of the asymptotic distribution of χ_P^2 (for $n \to \infty$ with N fixed). Pearson himself found this asymptotic distribution, namely the χ^2-distribution with $N - 1$ degrees of freedom. In studying the problem of a "best" choice of the set $\{I_j\}$ of intervals, Mann and Wald [17] adopted a non-parametric treatment, with $\nu = 2$ for the class Ω_ν above.

Another test not depending on a choice of intervals I_j could be made by using confidence belts for F as described in section 9 and rejecting H at the α level of significance if the graph of F_0 is not covered by the belt with confidence coefficient $1 - \alpha$.

The problem of randomness is usually non-parametric; in the univariate case the class ω of this problem is identical with the class Ω of the preceding The index ν and the class Ω for the problem of randomness would depend on the specific situation in which it arises. With two exceptions [42, 52], all tests of randomness proposed thus far have been functions of runs in the sample Two kinds of runs have been considered, runs up and down, and runs above and below the median [1, 4, 14, 19, 32, 44, 51] We note that the set S of permutations determined by ω is the set of all $n!$ permutations on the n coordinates of E. Suppose now $\nu = 2$. The proof [31] that all similar regions w have the randomization structure applies to this problem. On the other hand such a region w has the property $Pr\{\mathbf{E} \in w \mid F_n\} = \alpha$ for any F_n which is completely symmetrical in the coordinates Difficulty (i) discussed at the end of section 2 now arises if Ω contains such symmetrical alternatives. The definition of an appropriate class $\Omega - \omega$ of alternatives and the question of the power of tests against the alternatives make the problem of randomness a difficult one Beyond these few remarks we refer the reader to an expository paper by Wolfowitz [51] devoted to the problem in the previous issue of this journal, and to a paper by Wald and Wolfowitz [42] in the present issue. The latter paper is one of the exceptions, previously mentioned, not based on the method of ranks.

4. The problem of two samples. Suppose $X_1, \cdots, X_{m_1}$ and $Y_1, \cdots, Y_{m_2}$ are samples from univariate populations with c.d.f's $F(x)$ and $G(x)$ respectively, where we assume $F, G \in \Omega_\nu$, and that we wish to test the hypothesis that $F = G$. Write $Y_i = X_{i+m_1}$, so that with $n = m_1 + m_2$ we have

$$\Omega = \left\{F_n \mid F_n = \prod_{i=1}^{m_1} F(x_i) \cdot \prod_{j=m_1+1}^{n} G(x_j);\ F, G \in \Omega_\nu\right\},$$

$$\omega = \left\{F_n \mid F_n = \prod_{i=1}^{n} F(x_i);\ F \in \Omega_\nu\right\}.$$

The set S of permutations determining the subpopulation $\{E'\}$ consists of all $n!$ permutations on the n coordinates of E. The writer has shown [31] that no

similar regions exist in this case if $\nu = 0$, while if $\nu = 2$, 3, or 4 a similar region necessarily has the randomization structure.

The first non-parametric attack on this problem was given [26] by K. Pearson. The x-axis is divided up into intervals $I_1, \cdots, I_N$ as in section 3. Let m_{j1} and m_{j2} be the number of measurements from the first and second samples, respectively, falling in I_j, so that $\sum_{j=1}^{N} m_{jk} = m_k$, $k = 1, 2$. The statistic $T(E)$ used is

$$\chi^2_{P'} = (m_1 m_2)^{-1} \sum_{j=1}^{N} (m_1 m_{j2} - m_2 m_{j1})^2/(m_{j1} + m_{j2}),$$

with large values significant. In view of the remarks at the end of the last paragraph it would be necessary to calculate the distribution of $\chi^2_{P'}$ over the subpopulation $\{E'\}$ in order to get a similar region. Pearson found the asymptotic distribution of $\chi^2_{P'}$ under the null hypothesis to be the χ^2-distribution with $N - 1$ degrees of freedom.

A solution based on the method of randomization was proposed by Pitman [27]; the special case of this solution for $m_1 = m_2$ was published a little earlier by R. A. Fisher [6]. Pitman employed the numerical value of the difference of the sample means as statistic,

$$T(E) = \left| \sum_{i=1}^{m_1} x_i/m_1 - \sum_{j=m_1+1}^{n} x_j/m_2 \right|,$$

large values being significant. He fitted an incomplete Beta-distribution to the subpopulation distribution of his $T(E)$, and noted that this approximation gave a result identical with the usual t-test valid when the population distributions $F(x)$ and $G(x)$ are assumed normal with equal variances.

Turning now to tests based on the method of ranks, we mention here that one for the case $m_1 = m_2$ was given by R. A. Fisher as early as 1925, namely the "sign test" or "binomial series test" [3]. We may (and Fisher did) regard this as a test of a less restrictive hypothesis, and shall describe it in section 6. Between 1938 and 1940 several tests employing ranks were proposed for the problem of two samples. The earliest of these, by W. R. Thompson [36], was shown to be inconsistent (section 11) with respect to certain alternatives $F_n \in \Omega - \omega$ by Wald and Wolfowitz [40]. These authors used as statistic $U(R)$ the total number of runs in a sequence V of n elements constructed as follows: Rank the measurements of the combined sample in order of increasing magnitude. According as the j-th measurement in this rank order is from the first or second sample, put the j-th element of the sequence V equal to 1 or 2. In this test small values of the statistic $U(R)$ are regarded as significant. The test is now quite practicable (for $\nu = 2$) for certain ranges of m_1 and m_2. For m_1 and m_2 both ≤ 20, tables by Swed and Eisenhart [34] give the 1% and 5% significant values of $U(R)$. Wald and Wolfowitz proved that for $n \to \infty$ with $k = m_1/m_2$ fixed, the distribution of $U(R)$ is asymptotically normal with mean $2m_1/(1 + k)$ and variance $4km_1/(1 + k)^3$. Swed and Eisenhart computed that for $m_1 = m_2$ this

gives a very satisfactory approximation outside the range of their tables However, further computation needs to be done on the accuracy of this approximation for $m_1 \neq m_2$ and one of them >20

Another test based on ranks was advanced by Dixon [2], using as statistic $U(R)$ the random variable

$$C^2 = \sum_{j=1}^{m_1+1} [(m_1 + 1)^{-1} - n_j/m_2]^2,$$

where the integers n_j are defined thus: Let $Z_1 \leq Z_2 \leq \cdots \leq Z_{m_1}$ denote the measurements of the first sample arranged in rank order. Then n_j is the number of measurements in the second sample falling in the interval (Z_{j-1}, Z_j), where we define $Z_0 = -\infty$, $Z_{m_1+1} = +\infty$. Large values of C^2 are significant. Dixon tabulated the 1%, 5%, and 10% significant values of C^2 for m_1, $m_2 = 2, 3, \cdots, 10$; for larger m_1, m_2 he fitted a χ^2-distribution.

A paper by Smirnoff [33, 16] suggests the following as a statistic $U(R)$: Let $G_{m_1}(x)$ and $G_{m_2}(x)$ be the "empirical distribution functions" of the first and second samples, that is, $m_i G_{m_i}(x)$ is the number of measurements in the i-th sample $\leq x$ $(i = 1, 2)$ and take[7]

$$U(R) = (m_1^{-1} + m_2^{-1})^{-1} \sup_x |G_{m_1}(x) - G_{m_2}(x)|$$

with large values significant. Smirnoff showed that if $\nu = 2$ the asymptotic distribution of his statistic $U(R)$ is a certain c.d.f. $\Phi(\lambda)$, previously introduced by Kolmogoroff [15]. More specifically, let $\Phi_{m_1,m_2}(\lambda) = Pr\{U(R) \leq \lambda; \nu = 2\}$. Then if $n \to \infty$ with m_1/m_2 fixed, we have $\Phi_{m_1,m_2}(\lambda) \to \Phi(\lambda)$. The definition of $\Phi(\lambda)$ and references to tables of $\Phi(\lambda)$ are given in section 9. If instead of assuming $\nu = 2$ we take $\nu = 0$, the risk of type I errors may be controlled to the extent that $Pr\{\text{rejecting } H\} \leq \alpha$ for all $F_n \in \omega$, by employing Smirnoff's theorem stating $Pr\{U(R) \leq \lambda; \nu = 0\} \leq \Phi_{m_1,m_2}(\lambda)$, where $\Phi_{m_1,m_2}(\lambda)$ is defined above.

A test for the problem of two samples obtained by Wolfowitz by a modification of the likelihood ratio procedure will be discussed in section 12. When $m_1 = m_2$ the non-parametric analysis of variance tests of the "randomized blocks" type described in section 6 might also be used to test the more restricted hypothesis considered in this section.

The non-parametric problem of k samples has been attacked by Welch [46], who used the method of randomization with the analysis of variance ratio as statistic $T(E)$, and by Wolfowitz [50] with his modified likelihood ratio method.

In this as in all the other sections where several solutions of the same problem of statistical inference are described, the question as to the relative merits of the various solutions arises, and in every case the question is as yet mostly or entirely unanswered. The only easy conclusion about the tests of this section would seem to be that the tests of K. Pearson and Pitman are not consistent with

[7] We use the notations *sup* and *inf* respectively for least upper bound and greatest lower bound.

respect to certain subclasses of the admissible alternatives, according to the definition of section 11.

5. Independence. The classes Ω and ω defining the problem of independence have already been stated in section 2, in which we described Pitman's test [28] based on the randomization method and the use of $|r|$ as statistic $T(E)$, where r is the sample value of the Pearsonian correlation coefficient. Pitman fitted an incomplete Beta-distribution to the subpopulation distribution of r^2 and found the resulting approximation for $\nu = 2$ equivalent to the usual test employing the t-distribution and valid for the case of normality.

In section 2 we also mentioned the test earlier proposed by Hotelling and Pabst [9], which is based on the method ranks and employs the statistic $U(R) = |r'|$, where r' is the Spearman rank correlation coefficient. They proved that for $\nu = 2$ the distribution of r' is asymptotically normal if $F_n \in \omega$. Pitman's fitting of an incomplete Beta-distribution applies also to $(r')^2$, and Kendall, Kendall, and Smith [12] made numerical calculations indicating that this gives a better approximation than the normal distribution. Since r' is calculated from Σd^2, the sum of the squared rank differences, the latter may equivalently be used as the statistic $U(R)$, small and large values of Σd^2 being now both significant. Kendall, Kendall, and Smith [12] found the exact distribution of Σd^2 for the number of pairs $m \leq 8$. This work was anticipated by Olds [23], who calculated the exact distribution of Σd^2 for $m \leq 7$, and by fitting certain distributions for $m > 7$, gave a very useful table of the 1%, 2%, 4%, 10% and 20% significant values of Σd^2 for $m \leq 30$. It would be desirable to have these tables extended by inclusion of the 5% values.

M. G. Kendall [10] proposed another measure of rank correlation whose significant values are easier to calculate than those of Σd^2, but since the Olds' tables for the latter are available, Kendall's innovation does not seem to possess much practical advantage. Wolfowitz [50], using his modified likelihood ratio method, gave another test for independence and generalized it to the problem of independence of k random variables.

6. Analysis of variance. We suppose that we have $n = rc$ measurements arranged in a rectangular layout of r rows and c columns The r rows might correspond to the blocks and the c columns to the varieties in an agricultural experiment. The null hypothesis H is that of "no difference" in the column effects. The measurement in the i-th row and j-th column is supposed to be on a random variable[8] X_{ij} with c.d.f. $F^{(ij)}(x) = Pr\{X_{ij} \leq x\}$. Let us assume at first that all the X_{ij} are independent. The joint c.d.f. of the random variables $X_{1j}, \cdots, X_{rj}$ of the j-th column is then

$$F^{(j)}(x_1, \cdots, x_r) = Pr\{x_{1j} \leq x_1, \cdots, x_{rj} \leq x_r\} = \prod_{i=1}^{r} F^{(ij)}(x_i).$$

[8] The double subscript notation is more convenient here than that used in section 2, after the class ω has been defined the reader will see that the numbers n_i used in section 2 to describe the symmetry of the $F_n \in \omega$ are all equal to c, and the X_{ij} of the present section coincides with the X_{p_i+j} of section 2.

The symbol F_n for the joint c.d.f. of all n random variables now denotes $F_n(x_{11}, \cdots, x_{1c}; \cdots; x_{r1}, \cdots, x_{rc})$. Ω is the class of all F_n of the form

$$F_n = \prod_{j=1}^{c} F^{(j)}(x_{1j}, \cdots, x_{rj}),$$

where $F^{(j)}$ is defined by the preceding equation, and all $F^{(j)}$ are in a given class Ω_r. The hypothesis H states that the column distributions are all the same,

$$F^{(j)}(x_1, \cdots, x_r) = F^{(1)}(x_1, \cdots, x_r) \qquad (j = 2, 3, \cdots, c),$$

without specifying $F^{(1)}$. ω is thus the subclass of Ω comprising all F_n of the form

$$F_n = \prod_{j=1}^{c} F^{(1)}(x_{1j}, \cdots, x_{rj}).$$

The F_n in ω may be written

$$F_n = \prod_{i=1}^{r} \left\{ \prod_{j=1}^{c} F^{(1)}(x_{ij}) \right\}.$$

Regarding the factor in braces for fixed i, we see that it is left unchanged by any permutation of the c coordinates $x_{i1}, \cdots, x_{ic}$. The set S of permutations is thus determined, and the subpopulation $\{E'\}$ consists of the $(c!)^r$ points obtained by permuting among themselves the first set of c coordinates, the second set of c coordinates, $\cdots$, the r-th set of c coordinates of $E = (x_{11}, \cdots, x_{1c}; \cdots; x_{r1}, \cdots, x_{rc})$.

The above argument leading to the subpopulation $\{E'\}$ of $(c!)^r$ points is based squarely on the assumed independence of the n random variables X_{ij}. Suppose now that the X_{ij} are not known to be independent, as may happen in agricultural experiments [24]. To make the discussion concrete suppose in the $r \times c$ layout we have been considering, the rows refer to blocks (of plots) and the columns to varieties, so that the random variable X_{ij} is the yield of the j-th variety on the i-th block. We owe to R. A. Fisher the method of including early in the experiment a random process which leads to the same "equally likely" subpopulation of points $\{E'\}$ obtained before in the case of independence. This physical process which he calls "randomization" then permits the construction of critical regions by the "method of randomization" in the sense we have been using the term.

To explain the experimental process of randomization we shall imagine another $r \times c$ layout and a random set of mappings of the two layouts onto each other. In each block there are c plots and we now assume these numbered from 1 to c, the numbering to be held fixed. The second layout refers to the plots; the rows again correspond to the blocks, but the columns now correspond to the number of the plot in the block, thus the i, j cell represents the j-th plot in the i-th block. Now consider all 1:1 correspondences or mappings between the two layouts so that the i-th row always maps onto the i-th row ($i = 1, \cdots, r$). There are $s = (c!)^r$ such mappings M_k ($k = 1, \cdots, s$). Suppose under the mapping M_k the i, t cell in the block-plot layout maps on the i, j_k cell of the block-variety

layout, where $j_k = j_k(i, t)$, and the i, j cell of the latter corresponds to the i, t_k cell of the former, $t_k = t_k(i, j)$. The physical randomization process consists of choosing the mapping M_k so that all s mappings have the same probability $1/s$ of being chosen. In other words, the randomized block pattern is selected in such a way that all the s possible patterns have equal probabilities of being adopted in the experiment. Now let $Y_{ij}^{(k)}$ be the yield of the i, t plot if the variety assigned to it by the k-th pattern is planted there, and let $G^{(k)}(y_{11}, \cdots, y_{rc}) = Pr\{\text{all } Y_{ij}^{(k)} \leq y_{ij}\}$ be the joint c.d.f. of the $Y_{ij}^{(k)}$. In calculating the c.d.f. F_n of the X_{ij} associated with the first layout we must take account of the random process by which it is mapped onto the second:

$$\begin{aligned} F_n(x_{11}, \cdots, x_{rc}) &= Pr\{\text{all } X_{ij} \leq x_{ij}\} \\ &= \sum_{k=1}^{s} Pr\{\text{all } X_{ij} = Y_{i,t_k(i,j)}^{(k)}\}\, Pr\{Y_{i,t_k(i,j)}^{(k)} \leq x_{ij}\} \\ &= \sum_{k=1}^{s} s^{-1} G^{(k)}(x_{1,t_k(1,1)}, \cdots, x_{r,t_k(r,c)}). \end{aligned}$$

Ω consists of all F_n of the above form with $G^{(k)}$ in a given class, say $\Omega_\nu^{(n)}$. The hypothesis H of "no difference" of varieties asserts that the yields of the plots do not depend on the varieties planted on them, that is, that all $G^{(k)}$ are the same, $G^{(k)} = G^{(1)}$, without specifying $G^{(1)}$. ω is the subclass of Ω whose members are of the form

$$F_n = s^{-1} \sum_{k=1}^{s} G^{(1)}(x_{1,t_k(1,1)}, \cdots, x_{r,t_k(r,c)}).$$

It is now seen that any permutation in the set S previously considered merely rearranges the terms of the above sum, so that F_n remains invariant, and we have the same subpopulation $\{E'\}$ as before.

It is to be understood henceforth that either the X_{ij} are known to be independent or else an experimental randomization has been carried out as described above, so that in either case the above set $\{E'\}$ of $(c!)^r$ points is the "equally likely" subpopulation.

The first application in the literature of the randomization method is found in R. A. Fisher's "sign test" or "binomial series test" [3] for the case of randomized blocks with two columns ($c = 2$). Let D_i be the difference $X_{i1} - X_{i2}$. The statistic used is a function of the ranking only, namely the number of $D_i > 0$, small and large values being significant. For $\nu = 2$ its distribution under the null hypothesis is the binomial distribution with the n and p of the usual notation equal respectively to r and $\frac{1}{2}$. This test may be regarded as the special case when $c = 2$ of Friedman's rank method for analysis of variance to be described below.

Fisher later [5] proposed another test for the case $c = 2$ not based on ranks, and employing as statistic $T(E)$ the absolute value of the mean of the D_i defined above, with large values significant. The exact distribution of this statistic is very laborious to calculate unless r is very small, and K. R. Nair [20] pointed

out that the use of the numerical value of the median of the D_i (or one of the two central values when r is even) had the advantage of a very easily calculated distribution (if $\nu = 2$) The latter may be regarded as a modification of the rank method, the method of ranks being applied not in the $2r$-dimensionsl sample space as described in section 2 but in the r-dimensional space of the differences D_i. Nair also showed that the distributions of the range and of the midpoint of the range of the D_i are very simple

From here on we consider the general case $c \geq 2$, but when we speak of distributions they will be understood to be for the case when the null hypothesis is true and $\nu = 2$. Welch [45] considered using as $T(E)$ the usual analysis of variance ratio appropriate to testing for "no difference" of column effects. He transformed this to another statistic and calculated two moments of its subpopulation distribution. The first moment always agrees with that obtained under "normal theory", that is for the case $X_{ij} = C_i + Z_{ij}$, where the C_i are constants and the Z_{ij} are independently normally distributed with the same variance and zero means, but the second moment depends on the subpopulation $\{E'\}$. Here the exact distribution of the statistic is of course in general much more tedious to calculate than in the previous case $c = 2$; an incomplete Beta-distribution was fitted by Welch. Welch anticipated Pitman [29] who obtained the same results and got besides the third and fourth moments of Welch's statistic.

The method of ranks was applied by Friedman [7] who employed as statistic $U(R)$ a quantity formed as follows: Rank each set of row entries X_{ij} (for fixed i) in ascending order of magnitude, and let r_{ij} be the rank of X_{ij}, so that $r_{i1}, \cdots, r_{ic}$ is a rearrangement of the integers $1, \cdots, c$. Let $\bar{r}_j$ be the mean rank of the j-th column, $\bar{r}_j = \sum_{i=1}^{r} r_{ij}/r$, and take for $U(R)$

$$U = C_{rc} \sum_{j=1}^{c} [\bar{r}_j - \mathfrak{E}(\bar{r}_j)]^2,$$

where C_{rc} is a certain constant, and $\mathfrak{E}(\bar{r}_j)$ is calculated under the null hypothesis. For Friedman's choice of C_{rc}, U may be rapidly computed from the equivalent formula

$$U = -3r(c+1) + 12 \sum_{j=1}^{c} \left(\sum_{i=1}^{r} r_{ij} \right)^2 \Big/ [rc(c+1)].$$

In his paper Friedman included a proof of Wilks' that U has asymptotically the χ^2-distribution with $c - 1$ degrees of freedom as $r \to \infty$. Kendall and Smith [13] fitted to a transform of U a Fisher z-distribution with continuity corrections, obtaining a better approximation for small r than the χ^2-distribution. Wallis [43] independently proposed the use of $\eta_r^2 = U/[r(c-1)]$ as statistic and called it the rank correlation ratio Friedman in a later paper [8] on the subject, using exact values he had calculated, together with the Kendall-Smith approximation, published tables[9] of the 1% and 5% significant values of U for $c = 3, 4, 5, 6,$

[9] In these tables our U, r, c are denoted respectively by χ_r^2, m, n.

7, and sufficiently many values of r so that for these c and any r the significant values of U are now easily available.

After the above lengthy discussion for the "randomized blocks" case of analysis of variance, it will perhaps suffice merely to mention that the "Latin square" case may be similarly attacked from the non-parametric point of view, and this has been considered by Welch [45], Pitman [29], and E. S. Pearson [24]. They have taken as the statistic the usual analysis of variance ratio, and the work of Welch and Pitman in calculating the first two moments of its subpopulation distribution is even more tedious than in the "randomized blocks" case.

Part II. Non-parametric Estimation

7. Classical results on point estimation. Throughout part II the symbol $\mathbf{E}$ will always denote a random sample $X_1, \cdots, X_n$ from a univariate population with c.d.f. $F(x)$, where F is an unknown member of a given class to be stated in each case. The c.d.f. of $\mathbf{E}$ is thus

$$F_n(x_1, \cdots, x_n) = \prod_{i=1}^{n} F(x_i).$$

The problems of estimation can be stated without reference to the class Ω of admissible F_n; Ω would be obvious in every case.

Let $\theta = \theta(F)$ be a real number determined by F (a functional of F) for F in a certain class of univariate c.d.f's. Thus θ might be the mean of the distribution, in which case θ would be defined for all F possessing a first moment. We shall not call θ a parameter in order to avoid confusion with the parametric case R. A. Fisher's criteria of unbiasedness and of consistency for point estimation carry over without change from the parametric case. A statistic $T(E)$ is said to be an unbiased estimate of θ if $\mathfrak{E}(T) = \theta$. Write $\mathbf{E} = \mathbf{E}_n$ and $T = T_n$ to emphasize the sample size n, and assume that the statistic $T_n(\mathbf{E}_n)$ is defined for all n (or all $n >$ some n_0). Then we define $T_n(\mathbf{E}_n)$ to be a consistent estimate of θ if it converges stochastically to θ, that is, if $Pr\{|\ T_n - \theta\ | > h\} \to 0$ as $n \to \infty$, for every $h > 0$.

In the present paragraph it will be convenient to symbolize the class of F for which the i-th (absolute) moment exists; we denote it by $\Omega_{(i)} (i = 1, 2, \cdots)$. It is known[10] that a sufficient condition for the stochastic convergence of the sample mean $\bar{x}$ to the population mean is that $F \in \Omega_{(1)}$. Hence for all $F \in \Omega_{(1)}$, $\bar{x}$ is a consistent estimate of the population mean; furthermore it is unbiased. If we apply this result to the random variable $Y = X^2$, we find that for all $F \in \Omega_{(2)}$, $\sum_{i=1}^{n} x_i^2/n$ is a consistent unbiased estimate of the second moment of F about the origin. Similar statements may be made for higher moments. For $F \in \Omega_{(2)}$ one may show further that with Q defined as $\sum_{i=1}^{n} (x_i - \bar{x})^2$, the statistics Q/n and $Q/(n-1)$ are consistent estimates of the population variance, and the latter is unbiased.

[10] See, for example, J. L. Doob, *Annals of Math. Stat.*, Vol. 6 (1935), p. 163

If there exists a number M such that $F(M) = \frac{1}{2}$, it is called the median of the distribution. The median $\tilde{x}$ of a sample of odd size is the central X_i when the X_i are arranged in order of magnitude; for a sample of even size we may take $\tilde{x}$ to be the average of the two central values. It may be shown[11] that $\tilde{x}$ is a consistent estimate of M for F in the subclass of Ω_3 for which the probability density function $f(x)$ is continuous at $x = M$ and $f(M) \neq 0$.

8. Confidence intervals for an unknown median, for the difference of medians. Arrange the sample in rank order and denote the result by $Z_1 \leq Z_2 \leq \cdots \leq Z_n$, where $Z_1, \cdots, Z_n$ is a rearrangement of $X_1, \cdots, X_n$. The joint distribution of the Z_i (or any subset of the Z_i) is well known [49] if $F(x)$ is restricted to Ω_4, which we now assume. From this distribution theory it is easy to show that for any positive integer $k < \frac{1}{2}n$, the probability that the random interval (Z_k, Z_{n-k+1}) cover the unknown population median M is

$$Pr\{Z_k \leq M \leq Z_{n-k+1}\} = 1 - 2I_{\frac{1}{2}}(n - k + 1, k),$$

where

$$I_x(p, q) = \int_0^x t^{p-1}(1 - t)^{q-1}\,dt \Big/ \int_0^1 t^{p-1}(1 - t)^{q-1}\,dt$$

is the incomplete Beta-distribution tabulated by K. Pearson. The practicability of estimating M by means of the above relation in the non-parametric case was noted first by W. R. Thompson [35]. It is not difficult to calculate tables giving, for various sample sizes n, the maximum k for which $Pr\{Z_k \leq M \leq Z_{n-k+1}\} \geq .95$ or .99. This has been done for $n = 6$ to 81 by K. R. Nair [21], who listed the maximum k as well as $n - k + 1$ and $I_{\frac{1}{2}}(n - k + 1, k)$, so that the exact confidence coefficient is available. Nair also gave asymptotic formulas which are very accurate for $n > 81$.

It is clear how confidence intervals for the difference $d = M_2 - M_1$ of the medians of two univariate populations with c.d.f's known only to be in Ω_4 might be obtained by combining two probability statements of the above kind: Let the desired confidence coefficient be $1 - \alpha$, and form confidence intervals of the above type for M_1 and M_2 with confidence coefficient $1 - \frac{1}{2}\alpha$; write them $Pr\{\underline{M}_i \leq M_i \leq \bar{M}_i\} \geq 1 - \frac{1}{2}\alpha$. Then $Pr\{\underline{M}_2 - \bar{M}_1 \leq d \leq \bar{M}_2 - \underline{M}_1\} \geq 1 - \alpha$. Solutions like this which are easily obtained by the combining method in many problems are in general not very efficient.

Some work of Pitman's [27] may be regarded as a solution of the problem of estimating the difference of medians (or other quantiles, or means) of two populations in a case essentially more restricted than the preceding, but more general than the corresponding parametric case in which the distributions are assumed to differ only in location. To describe the nature of Pitman's result,

[11] This follows from the asymptotic distribution of $\tilde{x}$. See, for instance, [49], and combine section 4.53 with Theorem (A), p. 134.

let us revert to the notation introduced at the beginning of section 4, but add to the assumption that F and G are in a known class Ω_2 the *restrictive assumption* that F and G differ only in location, that is, that $G(x) = F(x - d)$ The problem is the interval estimation of the unknown constant d. Define the random variables $Z_i = Y_i - d$. After noting that the $m_1 + m_2$ random variables $X_1, \cdots, X_{m_1}, Z_1, \cdots, Z_{m_2}$ are all independently distributed with the same c.d.f F, Pitman was able to apply his results for the problem of two samples to show how functions $\underline{d}$ and $\bar{d}$ of $X_1, \cdot \cdot, X_{m_1}, Y_1, \cdots, Y_{m_2}$ could be calculated such that $Pr\{\underline{d} < d < \bar{d}\} \geq 1 - \alpha$ for $\nu = 0$, while for $\nu = 2$ the equality holds. After fitting an incomplete Beta-distribution Pitman found that the resulting approximate confidence intervals coincide with the well known ones employing the t-distribution and based on the assumption that F and G are normal with the same unknown variance.

9. Confidence limits for an unknown distribution function. Consider in an x, y-plane the graph g of the unknown c.d.f., g being the locus of the equation $y = F(x)$, and the possibility of covering g with random regions $\mathfrak{R}(\mathbf{E})$ depending on the sample $\mathbf{E}$. Wald and Wolfowitz [39] have shown how for given n and α it is possible in a large variety of ways to define regions $\mathfrak{R}(\mathbf{E})$ such that $Pr\{\mathfrak{R}(\mathbf{E}) \supset g\}$, the probability that the random region $\mathfrak{R}(\mathbf{E})$ cover the unknown graph g, is $1 - \alpha$ for all $F \in \Omega_2$. Instead of describing their general method we shall limit ourselves to a special case. This is a very neat solution the necessary distribution theory for which was developed earlier by Kolmogoroff [15].

Let $G_n(x)$ be the "empirical distribution function" of the sample: $nG_n(x)$ is the number of $X_i \leq x$ Define the random variable

$$D_n = \sqrt{n} \sup_x | F(x) - G_n(x) |,$$

and let $\Phi_n(\lambda)$ be the c.d.f. of D_n, $\Phi_n(\lambda) = Pr\{D_n \leq \lambda\}$. Kolmogoroff proved that $\Phi_n(\lambda)$ is independent of $F \in \Omega_2$, and that as $n \to \infty$, $\Phi_n(\lambda) \to \Phi(\lambda)$ uniformly in λ, where $\Phi(\lambda)$ is defined by the rapidly converging Dirichlet series

$$\Phi(\lambda) = \sum_{k=-\infty}^{+\infty} (-1)^k \exp(-2k^2\lambda^2).$$

A small table of values of the function $\Phi(\lambda)$ was given by Kolmogoroff [15], and a larger one by Smirnoff [33]. Define $\lambda_{n,\alpha}$ from $\Phi_n(\lambda_{n,\alpha}) = 1 - \alpha$, and λ_α from $\Phi(\lambda_\alpha) = 1 - \alpha$. Values of λ_α for $\alpha = .05, .02, .01, .005, .002, .001$ were listed by Kolmogoroff [16]. Now $1 - \alpha$ is the probability that

$$\sqrt{n} \sup_x | F(x) - G_n(x) | \leq \lambda_{n,\alpha}$$

if $F \in \Omega_2$. The above inequality is equivalent to

$$G_n(x) - \lambda_{n,\alpha}/\sqrt{n} \leq F(x) \leq G_n(x) + \lambda_{n,\alpha}/\sqrt{n} \qquad \text{(all } x\text{)}.$$

If we take as $\mathfrak{R}(\mathbf{E})$ the intersection of the region between the graphs of the functions $G_n(x) \pm \lambda_{n,\alpha}/\sqrt{n}$, with the strip $0 \leq y \leq 1$, we have $Pr\{\mathfrak{R}(\mathbf{E}) \supset g\} =$

$1 - \alpha$. The values of $\lambda_{n,\alpha}$ have not been tabulated, but for practical purposes of determining an unknown c.d.f one would usually require a large n, and the tabulated values of λ_α could then be used.

With $\Phi_n(\lambda)$ defined as the c d f. of D_n for $F \in \Omega_2$, Kolmogoroff has shown further that for $F \in \Omega_0$, $Pr\{D_n \leq \lambda\} \geq \Phi_n(\lambda)$. This gives the beautiful result that the above confidence belt is valid in the most general case where $F \in \Omega_0$, in the sense that for the above defined $\mathfrak{R}(\mathbf{E})$, $Pr\{\mathfrak{R}(\mathbf{E}) \supset g\} \geq 1 - \alpha$

10. Tolerance limits. An ingenious formulation and solution of a non-parametric estimation problem was given by Wilks [47]. Let us say that an interval (x', x'') covers a proportion π of a population with c d.f. $F(x)$ if $F(x'') - F(x') = \pi$. In the notation of section 8, Wilks considered the proportion B covered by the interval (Z_k, Z_{n-m+1}) extending from the k-th smallest observation to the m-th largest, $B = F(Z_{n-m+1}) - F(Z_k)$. B is a random variable depending on the sample but is *not* a statistic since it depends also on the unknown c.d.f. $F(x)$. However, Wilks noted that the c.d.f $G(b)$ of B is independent of $F \in \Omega_4$, in fact, for $0 < b < 1$,

$$G(b) = I_b(n - k - m + 1, k + m),$$

where $I_x(p, q)$ is defined in section 8 After k, m, a fixed proportion b, and a confidence coefficient $1 - \alpha$ have been chosen, the equation $G(b) = \alpha$ determines the sample size n for which we can then make the following assertion without any knowledge of F except that $F \in \Omega_4$: The probability is $1 - \alpha$ that in a sample size n the random interval (Z_k, Z_{n-m+1}) will cover at least 100 b% of the population.[12]

Wilks considered, among other extensions of his method, tolerance limits for multivariate distributions in which the variables are known to be independent, and the estimation of proportions in a second sample (instead of in the population) on the basis of a first sample [48]. The latter problem involves the calculation of $P(b; n, N, k, m)$, the probability that if a first sample of n is taken and then a second sample of N, a proportion b or more of the second sample will lie in the interval (Z_k, Z_{n-m+1}) determined from the first sample. Wilks' derivation of P requires the assumption that $F \in \Omega_4$, but a simple auxiliary argument (related to the method of randomization by ranks) will extend the validity to the case $F \in \Omega_2$: The complete set of $n + N$ variates is independently distributed, each with the same c.d.f. $F \in \Omega_2$. All $(n + N)!$ possible rankings (excluding the "tied" ranking R_0) as defined in section 2 then have the same probability $1/(n + N)!$ The fraction of these rankings for which the statement about proportions in the second sample is correct is a function of b, n, N, k, m only, and not of $F \in \Omega_2$, and this fraction is the desired P. Since P is the same for all $F \in \Omega_2$ it must of course coincide with the value calculated by Wilks for $F \in \Omega_4$. It would be desirable for practical purposes to extend the validity of the tolerance

[12] For fixed b, $G(b)$ of course takes on discrete values with n, so one would either choose the n giving $G(b)$ the nearest value to α or else the greatest value $\leq \alpha$.

limits of the first paragraph, concerning proportions in the population, at least to the case $F \in \Omega_3$. The extension to Ω_2 would follow immediately if the intuitively reasonable statement $1 - G(b) = \lim_{N\to\infty} P(b; n, N, k, m)$ could be justified for $F \in \Omega_2$.

The multivariate case when independence is not assumed was successfully attacked by Wald [38]. We shall describe here his solution for the bivariate case: Let (X_i, Y_i), $i = 1, \cdots, n$, be a sample from a population with bivariate c.d.f. $F(x, y) \in \Omega_4^{(2)}$, that is, F is of the form

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(\xi, \eta)\, d\eta\, d\xi,$$

where $f(x, y)$ is continuous, but otherwise unknown. Plot the points (X_i, Y_i) in an x, y-plane and choose four (small) integers k_1, m_1, k_2, m_2. Draw vertical lines (parallel to the y-axis) passing through the points with the k_1-th smallest and m_1-th largest abscissas. Considering *only* the $n - k_1 - m_1$ points inside these vertical lines (the probability of equal abscissas is zero), draw two horizontal lines passing through the points with k_2-th smallest and m_2-th largest ordinates. Let J be the rectangle bounded by the four lines and consider the proportion B of the population covered by the rectangle, $B = \int_J dF(x, y)$. Then the c.d.f. $G(b)$ of B is given by the previous formula in terms of the incomplete Beta-distribution with $k + m = k_1 + k_2 + m_1 + m_2$, and is thus independent of $f(x, y)$. Choose k_1, k_2, m_1, m_2, b, and α. Then the equation $G(b) = \alpha$ determines the sample size n for which the probability is $1 - \alpha$ that the random rectangle J will cover at least 100 b% of the population. Wald showed further how a series of rectangles instead of a single rectangle might advantageously be used in the case of highly correlated X, Y.

It would be most useful to have tables of n corresponding to $\alpha = .05$ and .01, some values of b close to unity, and a few small values of $k + m$, say, $k + m = 2, 4, \cdots, 2r$. The table could then be used for the univariate, bivariate, $\cdots$, r-variate cases with various choices of k_j, m_j, such that $\Sigma(k_j + m_j) = k + m$. Entries for $k + m = 4$ have been given by Wald [38, p. 55].

Part III. Toward a General Theory

11. The criterion of consistency. All the concepts of Part III have been carried over from, or suggested by, corresponding ones earlier developed for the parametric theory. Consistency of point estimation was defined in section 7. Wald and Wolfowitz [40] have generalized the notion of consistency to tests so that it is applicable in the non-parametric case. We have heretofore specified the hypothesis H and its admissible alternatives by means of classes of n-variate c.d.f's F_n. We now assume that H and its admissible alternatives can be framed as statements about one or more populations, independent of n. Thus in the problem of two samples (section 4) H may be taken as the statement that the c.d.f's F and G of the two populations are the same member of Ω_r, while the

admissible alternatives are statements that F and G are any two different members of Ω_r. Returning to the general case, we assume that a sequence of tests is under consideration, say, $\mathfrak{T}_1, \mathfrak{T}_2, \cdots$, such that as $j \to \infty$, the size of the sample in $\mathfrak{T}_j$ from each of the populations becomes infinite. The sequence $\{\mathfrak{T}_j\}$ may be called simply a "test" and is said to be consistent if the probability of rejection of H by $\mathfrak{T}_j$ approaches unity as $j \to \infty$ whenever an admissible alternative to H is true It has been suggested [50] that consistency is a minimal requirement for a good test. In order to allow for the analogue of the "common best critical regions" in the parametric theory,[13] it would be better to define consistency with respect to any given subset of the admissible alternatives and then require consistency with respect to the subset appropriate to the specific situation in which the test is to be used.

Wald and Wolfowitz [40] proved that under certain restrictions on the admissible F, G in the problem of two samples their test based on runs (section 4) is consistent, while another previously proposed test is not. Judging from their work, we may expect that, while inconsistency proofs may be easy, consistency proofs will be difficult.

12. Likelihood ratio tests. A definition of the Neyman-Pearson likelihood ratio criterion[14] λ for testing the hypothesis H (we use the notation of section 2), which would yield the usual result in the parametric case, would be the following: Let $C(E;\delta)$ be a cube of edge 2δ in the sample space W with center at the point E and faces parallel to the coordinate hyperplanes, and let $P(E,\delta \mid F_n)$ be the "probability put into the cube by the c d f. F_n", that is, $P(E,\delta \mid F_n) = \int_{C(E;\delta)} dF_n$. Define

$$\lambda(E, \delta) = [\sup_{F_n \in \omega} P(E, \delta \mid F_n)]/[\sup_{F_n \in \Omega} P(E; \delta \mid F_n)],$$

$$\lambda = \lambda(E) = \lim_{\delta \to 0} \lambda(E; \delta).$$

This definition of λ is not useful in the non-parametric case as λ turns out in general to be independent of E, the reader may easily verify this for the problem of two samples (section 4)

Having seen now that the likelihood ratio does not carry over to the non-parametric case in an obvious way, we are in a position to appreciate a bold stroke by Wolfowitz [50]. He begins by limiting the critical regions to be considered to the relatively small class obtainable by the method of ranks (section 2) Let $R = R(\mathbf{E})$ be the ranking of the sample point $\mathbf{E}$, so that the random variable R takes on the possible values $R_0, R_1, \cdots, R_s$, and let $P(R_\sigma \mid F_n) = Pr\{R = R_\sigma \mid F_n\}$.

[13] J Neyman and E S Pearson, "On the problem of the most efficient tests of statistical hypotheses", *Phil. Trans. Roy. Soc London*, A, Vol. 231 (1933), pp 289-337.

[14] J. Neyman and E. S Pearson, *Biometrika*, Vol 20A (1928), p 264

Then Wolfowitz takes the likelihood ratio to be the following function of the ranking R:

$$\Lambda(R) = [\sup_{F_n \in \omega} P(R \mid F_n)]/[\sup_{F_n \in \Omega} P(R \mid F_n)].$$

His modified likelihood ratio test then consists of applying the method of ranks (section 2) with $\Lambda(R)$ as the statistic, small values being regarded as significant. If Ω is a class of continuous F_n, all rankings $R \neq R_0$ have the same probability $1/s$ under the null hypothesis, while $P(R_0 \mid F_n) = 0$ for all $F_n \in \Omega$. Then the numerator of $\Lambda(R)$ is $1/s$, and we may thus use the denominator of $\Lambda(R)$ as statistic with large values significant. Wolfowitz' modification has one advantage we don't always find with the usual parametric method: it always leads to similar regions since it is a special case of the randomization method.

In applying his method to examples Wolfowitz finds it necessary to resort each time to an approximation in calculating his statistic $\Lambda(R)$. Instead of taking the "sup" over Ω as in the definition, he takes it instead over a subclass Ω' of Ω which lends itself more easily to calculation Thus in the problem of two samples with $\nu = 2$, whereas Ω is the class defined in section 4 with F, G in Ω_2, the class Ω' is the subclass of Ω obtained by further limiting F, G as follows: The x-axis is divided up into a number of disjoint intervals, equal to the total number of runs in the sequence V defined in connection with the Wald-Wolfowitz test in section 4. If the j-th run in V is a run of 1's the restriction $G(x) = 0$ in the j-th interval is imposed, if the j-th run is a run of 2's, $F(x) = 0$ in the j-th interval. The intervals in which F, G are permitted to assign positive probability then correspond in order and number to the two kinds of runs. With this restriction the (twice) modified likelihood ratio statistic is found to be

$$\sum_i \sum_j (l_{ij} \log l_{ij} - \log l_{ij}!),$$

where l_{ij} is the number of elements in the j-th run of i's ($i = 1, 2$). Large values are significant. For large samples the asymptotic distribution of the statistic falls out as a special case of a general theorem of Wolfowitz.

In the same paper Wolfowitz obtained modified likelihood ratio tests for the problem of k samples and the problem of independence of two or more random variables.

In his examples the author states that the maximizing F_n in Ω' is "essentially the same" as the maximizing F_n in Ω, at least for the significant rankings R_σ and for large samples. The necessity of this approximation procedure is somewhat disturbing, as is the restriction to the method of ranks. Since it does not seem possible to give a definition of likelihood ratio tests sufficiently broad to include the non-parametric case, yet yielding the usual result in the parametric case, we are denied even the small comfort of saying that at least in special cases the method is known to yield optimum results. In some problems the set $\{R_\sigma\}$ of rankings, corresponding to the set $\{w_\sigma\}$ of regions in W which serves to separate the s points of the subpopulations $\{E'\}$ defined in section 2, is not

unique—consider for instance the problem of two samples when the populations are bivariate—and in such cases the method as defined above would not give a unique result. These remarks are intended to point the need for further investigation and cannot detract from the ingenuity of the method—the first general process that has been suggested for choosing one out of the welter of similar regions yielded by the randomization method.

13. Wald's formulation of the general problem of statistical inference. A formulation of the general problem of statistical inference broad enough to cover the non-parametric case, and including estimation and tests as well as statistical problems classifiable under neither of these headings, has been given by Wald [37]. This formulation extends certain concepts he had applied earlier[15] to the parametric case.

In this last section we shall permit ourselves a somewhat more abstract terminology and notation than before. As in section 2, $\mathbf{E} = (X_1, \cdots, X_n)$ will denote the sample; $F_n(E)$, its c.d.f.; W, the n-dimensional Euclidean space of E, the sample space; and Ω, the space of admissible F_n. Of central importance is a given class $\mathfrak{S}$ appropriate to the problem, $\mathfrak{S} = \{\omega_\beta\}$, whose members ω_β are (not necessarily disjoint) subsets of Ω, $\Omega = \bigcup_\beta \omega_\beta$. To every $\omega_\beta \in \mathfrak{S}$ there corresponds a hypothesis $H(\omega_\beta): F_n \in \omega_\beta$, so that there is a 1:1 correspondence between the members of the set $\mathfrak{S}$ and those of the set $\{H(\omega_\beta)\}$ of hypotheses. The general problem of statistical inference, according to Wald, is the choice of a decision function $\Delta(E)$ mapping W into $\mathfrak{S}$. For every $E \in W$ a decision function $\Delta(E)$ uniquely selects an element ω_β of $\mathfrak{S}$, $\omega_\beta = \Delta(E)$. Its statistical import is that when the sample point $\mathbf{E}$ equals E, we agree to accept the hypothesis $H(\omega_\beta)$ determined by $\Delta(E) = \omega_\beta$.

Before introducing any further definitions let us illustrate the preceding ones. In any problem of testing a hypothesis, the set $\mathfrak{S}$ has just two members ω_1 and ω_2 which we have heretofore denoted by ω and $\Omega - \omega$, respectively. The decision function $\Delta(E)$ then takes on just these two values, in fact, $\Delta(E) = \omega_2$ for E in the critical region w of the test, and $\Delta(E) = \omega_1$ for $E \in W - w$.

To illustrate the definitions in the case of point estimation, consider estimating the median M of a univariate population with c.d.f. $F(x)$. Ω would be the class of F_n of the form $\prod_{i=1}^n F(x_i)$ with, say, $F \in \Omega_4$ and $F'(M) \neq 0$ (which is sufficient to insure a unique M). The index β could now be identified with M, so that its domain is the real line, and $\omega_\beta = \{F_n \mid M(F) = \beta\}$. The classes ω_β would be disjoint in this case and each would contain an infinite number of F_n. The problem of estimating the unknown M may be said to be the choice of a decision function $\Delta(E)$: When $\mathbf{E} = E$ we accept $H(\omega_\beta): F_n \in \omega_\beta = \Delta(E)$, meaning in this case simply that we accept the statement that M equals the β determined by $\Delta(E)$.

---

[15] A. Wald, "Contributions to the theory of statistical estimation and testing hypotheses", *Annals of Math Stat.*, Vol. 10 (1939), pp. 299-326

Suppose next that instead of the point estimation of M just discussed we are interested in the interval estimation of M. We define Ω as above, and now take the index β to consist of a pair a, b of real numbers. An interval estimate $a \leq M \leq b$ may be regarded as an acceptance of the hypothesis $H(\omega_{a,b}): F_n \in \omega_{a,b}$, where $\omega_{a,b}$ is the subclass of Ω consisting of all F_n for which $M(F)$ lies in the interval $a \leq M \leq b$. The set $\mathfrak{S}$ now consists of all classes $\omega_{a,b}$ with $-\infty < a < b < +\infty$. Here as in the general case of interval estimation the classes ω_β of the set $\mathfrak{S}$ are not disjoint. The decision function $\Delta(E)$ adopted in section 8 is $\Delta(E) = \omega_{a,b}$ with $a = z_k$, $b = z_{n-k+1}$, where $z_1 \leq z_2 \leq \cdots \leq z_n$ is a rearrangement of the coordinates $x_1, \cdots, x_n$ of E.

An example of a problem neither of estimation nor testing would be the following: Let Ω be as above. Two real numbers A and B $(A < B)$ are given and it is required to decide on the basis of the sample $\mathbf{E}$ to which of the three classes $-\infty < M < A$, $A \leq M \leq B$, $B < M < +\infty$ the unknown median M belongs. Here the set $\mathfrak{S}$ would consist of three disjoint classes ω_1, ω_2, ω_3: where ω_1 is the subclass of Ω consisting of F_n with $M(F) < A$, etc.

We return now to the general case. Before defining a "best" decision function $\Delta = \Delta^*$, Wald asks that there be a given weight function $w(F_n, \omega_\beta)$ defined on the product space $\Omega \times \mathfrak{S}$. The weight function $w(F_n, \omega_\beta)$ is a real-valued function evaluating the loss involved in accepting $H(\omega_\beta)$, the statement that the unknown c.d.f. of $\mathbf{E}$ is a member of ω_β, when the unknown c.d.f. is actually F_n. If $F_n \in \omega_\beta$ we make no error in accepting $H(\omega_\beta)$, and in this case w is defined to be zero. Its value otherwise is required to be non-negative. In this theory the choice of the weight function is regarded as essentially *not* a mathematical problem, but the choice is to stem out of the very specific situation in which the statistical inference is to be made. In an industrial problem w might be the financial loss incurred when a certain kind of error is made.

After w is given, the decision functions Δ are to be restricted to the class for which $w(F_n, \Delta(E))$ is a Borel-measurable function of E for all $F_n \in \Omega$; note that w depends on E only through Δ, not through F_n. The expected value of w for a particular F_n is called the risk function; it depends of course on the decision function Δ and the weight function w as well as on F_n. Denote it by

$$r(\Delta, w \mid F_n) = \int_W w(F_n, \Delta(E))\, dF_n(E).$$

Since the true F_n is unknown, so in general will be the true value of the risk function associated with a particular decision function Δ. We might call

$$r(\Delta, w) = \sup_{F_n \in \Omega} r(\Delta, w \mid F_n)$$

the maximum risk associated with the decision function Δ. Wald defines Δ^* to be the "best" decision function relative to the weight function w if the maximum risk $r(\Delta, w)$ is minimum for $\Delta = \Delta^*$. He points out that the "best" decision

function might be defined as one which minimizes some weighted mean, taken over all $F_n \in \Omega$, of the risk function $r(\Delta, w \mid F_n)$, but that the above definition of the "best" decision function has certain advantages Thus under certain restrictions on Ω and w, the risk function $r(\Delta^*, w \mid F_n)$ is independent of $F_n \in \Omega$, that is, we then know the exact value of the risk, regardless of what the true F_n may be. This is analogous to the desirable situations where confidence intervals are known, and the probability of a false statement (to the effect that the unknown quantity is in a given region when it is not) is then a constant independent of the unknown quantity.

Wald's theory is suggestive and formally very satisfying, but one would like to see some specific examples of its application to non-parametric cases. A discouraging aspect, not shared by the older Neyman-Pearson theory, lies in the very refinement that a decision function is declared best with respect to a very particular weight function w. An attractive possibility would be to impose a metric on Ω or on a related function space, and to let w be the distance function. In the problem of two samples for example, after metrizing Ω_2, the weight w assigned to accepting H might be taken as the distance between F and G in the notation of section 4. A suitable choice of metric might yield a weight function appropriate to a large variety of situations. The difficulties of finding a distance function which is intuitively satisfactory and analytically tractable in calculating the risk function are no doubt formidable. The device of metrizing a space of distribution functions was used by Mann and Wald in a different connection [17], but their choice of distance function, while appropriate to their problem, would not be satisfactory here

Also still lacking is any general theory relating the three concepts discussed in Part III. The following questions have been answered, at least for some specific examples, in the parametric case, but are still untouched in the non-parametric case: Are likelihood ratio tests consistent? Is there a simple weight function w relative to which the likelihood ratio test becomes a "best" test, or asymptotically a "best" test? If a test is "best" relative to a given weight function, with respect to what set of alternatives is it consistent?

In conclusion let us emphasize the need for *constructive* methods of obtaining "good" and "best" tests and estimates in the non-parametric case. Recalling the history of the parametric case we may judge that half the battle was the definition of "good" and "best" statistical inference. Progress in the non-parametric case has been made in the direction of definition, mainly by carrying over or modifying criteria originally advanced for the parametric case. However, besides criteria for "good" and "best" tests and estimates, we have in the parametric case a large body of constructive theory which may be applied in particular examples to yield the optimum tests or estimates; thus we have the Fisher theory of maximum likelihood statistics for point estimation, and the constructive theorems of the Neyman-Pearson theory for the existence of critical regions of types A, A_1, B, B_1, and the related types of "best" confidence inter-

vals. The contrasting lack of any general constructive methods[16] at present challenges us in the non-parametric theory.

BIBLIOGRAPHY OF STATISTICAL INFERENCE IN THE NON-PARAMETRIC CASE

[1] W E CAMPBELL, "Use of statistical control in corrosion and contact resistance studies", *Bell Telephone System Technical Publications*, Monograph B-1350 (1942).
[2] W. J. DIXON, "A criterion for testing the hypothesis that two samples are from the same population", *Annals of Math. Stat.*, Vol. 11 (1940), pp 199–204
[3] R A. FISHER, *Statistical Methods for Research Workers*, section 24, example 19, Oliver and Boyd, Edinburgh, 1925.
[4] R. A. FISHER, "On the random sequence", *Quart. J. Roy Meteor Soc.*, Vol 52 (1926), p. 250.
[5] R A. FISHER, *The Design of Experiments*, section 21, Oliver and Boyd, Edinburgh, 1935.
[6] R. A FISHER, "Coefficient of racial likeness and the future of craniometry", *J. R Anthrop Inst*, Vol 66 (1936), pp. 57–63
[7] M. FRIEDMAN, "The use of ranks to avoid the assumption of normality", *J. Amer. Stat. Ass.*, Vol. 32 (1937), pp 675–701
[8] M FRIEDMAN, "A comparison of alternative tests of significance for the problem of m rankings", *Annals of Math. Stat.*, Vol 11 (1940), pp. 86–92
[9] H. HOTELLING AND M. R. PABST, "Rank correlation and tests of significance involving no assumptions of normality", *Annals of Math Stat.*, Vol. 7 (1936), pp 29–43.
[10] M. G. KENDALL, "A new measure of rank correlation", *Biometrika*, Vol. 30 (1938), pp. 81–93.
[11] M. G. KENDALL, "Note on the estimation of a ranking", *J. Roy. Stat. Soc*, Vol. 105 (1942), pp. 119–121.
[12] M. G. KENDALL, S. F. H. KENDALL, AND B. B. SMITH, "The distribution of Spearman's coefficient of rank correlation in a universe in which all rankings occur an equal number of times", *Biometrika*, Vol. 30 (1939), pp. 251–273.
[13] M. G KENDALL AND B. B. SMITH, "The problem of m rankings", *Annals of Math. Stat.*, Vol. 10 (1939), pp. 275–287.
[14] W. O. KERMACK AND A G MCKENDRICK, "Tests for randomness in a series of observations", *Proc. Roy. Soc. Edinburgh*, Vol. 57 (1937), pp 228–240.
[15] A. KOLMOGOROFF, "Sulla determinatione empirica di una legge di distributione", *Giornale Ist. Ital. Attuari*, Vol. 4 (1933), pp 83–91.
[16] A. KOLMOGOROFF, "Confidence limits for an unknown distribution function", *Annals of Math. Stat*, Vol. 12 (1941), pp 461–463.
[17] H B MANN AND A. WALD, "On the choice of the number of class intervals in the application of the chi square test", *Annals of Math Stat.*, Vol. 13 (1942), pp. 306–317
[18] H C. MATHISEN, "A method of testing the hypothesis that two samples are from the same population", *Annals of Math Stat.*, Vol. 14 (1943), pp 188–194.
[19] F. MOSTELLER, "Note on application of runs to quality control charts", *Annals of Math Stat.*, Vol. 12 (1941), pp 228–232.
[20] K. R. NAIR, "The median in tests by randomization", *Sankhyā*, Vol. 4 (1940), pp. 543–550.
[21] K. R. NAIR, "Table of confidence interval for the median in samples from any continuous population", *Sankhyā*, Vol. 4 (1940), pp. 551–558 Second column heading on p. 556 should be changed from n to k

[16] Wolfowitz' modified likelihood ratio method is a general constructive method of getting tests which we hope to be "good", but until its optimum properties are investigated, his method does not constitute an exception to the thesis of this paragraph We remark however that the popularity of the usual parametric likelihood ratio method did not await Wald's recent proof of its optimum qualities (to appear in *Trans of the Amer. Math. Soc.*)

[22] J. NEYMAN, "Basic ideas and some recent results of the theory of testing statistical hypotheses", sections 20, 21, *J Roy. Stat Soc*, Vol. 105 (1942), pp 292-327

[23] E G. OLDS, "Distribution of sums of squares of rank differences for small numbers of individuals", *Annals of Math. Stat*, Vol 9 (1938), pp 133-148.

[24] E. S PEARSON, "Some aspects of the problem of randomization", *Biometrika*, Vol 29 (1937), pp 53-64 and Vol. 30 (1938), pp 159-179

[25] K. PEARSON, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling", *Phil Mag*, 5 ser., Vol 50 (1900), pp. 157-175

[26] K PEARSON, "On the probability that two independent distributions of frequency are really samples from the same population", *Biometrika*, Vol. 8 (1911), pp 250-254.

[27] E. J. G. PITMAN, "Significance tests which may be applied to samples from any populations", *Suppl. J Roy. Stat Soc.*, Vol. 4 (1937), pp 117-130

[28] E J G. PITMAN, "Significance tests which may be applied to samples from any populations II The correlation coefficient test" *Suppl. J. Roy Stat Soc*, Vol 4 (1937), pp 225-232

[29] E J G. PITMAN, "Significance tests which may be applied to samples from any populations. III. The analysis of variance test" *Biometrika*, Vol. 29 (1938), pp. 322-335

[30] S R. SAVUR, "The use of the median in tests of significance", *Proc. Ind Acad Sc.* (A), Vol 5 (1937), pp 564-576.

[31] H SCHEFFÉ, "On a measure problem arising in the theory of non-parametric tests", *Annals of Math. Stat*, Vol. 14 (1943), pp 227-233.

[32] W A. SHEWHART, "Contribution of statistics to the science of engineering", Univ. of Pa. Bicentennial Conference Volume on *Fluid Mechanics and Statistical Methods in Engineering*, pp. 97-124. Also *Bell Telephone System Technical Publications*, Monograph B-1319

[33] N SMIRNOFF, "On the estimation of the discrepency between empirical curves of distribution for two independent samples", *Bull Math. Univ. Moscou, Série Int*, Vol. 2, fasc. 2 (1939)

[34] F. S. SWED AND C EISENHART, "Tables for testing randomness of grouping in a sequence of alternatives", *Annals of Math Stat*, Vol 14 (1943), pp 66-87.

[35] W R. THOMPSON, "On confidence ranges for the median and other expectation distributions for populations of unknown distribution form", *Annals of Math. Stat.*, Vol. 7 (1936), pp. 122-128.

[36] W. R THOMPSON, "Biological applications of normal range and associated significance tests in ignorance of original distribution forms", *Annals of Math. Stat.*, Vol. 9 (1938), pp. 281-287

[37] A. WALD, *On the principles of statistical inference*, Notre Dame, Indiana, 1942

[38] A. WALD, "An extension of Wilks' method for setting tolerance limits", *Annals of Math. Stat*, Vol 14 (1943) pp 45-55.

[39] A WALD AND J WOLFOWITZ, "Confidence limits for continuous distribution functions", *Annals of Math Stat.*, Vol. 10 (1939), pp. 105-118.

[40] A WALD AND J WOLFOWITZ, "On a test whether two samples are from the same population", *Annals of Math. Stat*, Vol. 11 (1940), pp. 147-162.

[41] A. WALD AND J. WOLFOWITZ, "Note on confidence limits for continuous distribution functions", *Annals of Math. Stat*, Vol 12 (1941), pp. 118-119

[42] A WALD AND J WOLFOWITZ, "An exact test for randomness in the non-parametric case, based on serial correlation", *Annals of Math Stat.*, this issue

[43] W A WALLIS, "The correlation ratio for ranked data", *J Amer Stat. Ass.*, Vol. 34 (1939), pp 533-538

[44] W A. Wallis and G. H Moore, *A significance test for time series*, Technical paper 1 (1941), National Bureau of Economic Research, New York.

[45] B. L. Welch, "On the z-test in randomized blocks and Latin squares", *Biometrika*, Vol. 29 (1937), pp 21-52

[46] B L. Welch, "On tests of homogeneity", *Biometrika*, Vol. 30 (1938), pp 149-158.

[47] S. S Wilks, "Determination of sample sizes for setting tolerance limits", *Annals of Math Stat*, Vol 12 (1941), pp 91-96.

[48] S S. Wilks, "Statistical prediction with special reference to the problem of tolerance limits", *Annals of Math. Stat.*, Vol 13 (1942), pp 400-409.

[49] S S. Wilks, *Mathematical Statistics*, Princeton University Press, 1943

[50] J Wolfowitz, "Additive partition functions and a class of statistical hypotheses", *Annals of Math Stat*, Vol. 13 (1942), pp. 247-279.

[51] J Wolfowitz, "On the theory of runs with some applications to quality control", *Annals of Math. Stat.*, Vol. 14 (1943), pp. 280-288

[52] L C. Young, "On randomness in ordered sequences", *Annals of Math. Stat*, Vol. 12 (1941), pp. 293-300.

Useful Distribution Theory Not Included Above

[53] E. J. Gumbel, "Les valeurs extrêmes des distributions statistiques", *Annales de l'Inst. H. Poincaré*, Vol 5 (1935), pp 89-114.

[54] W. O Kermack and A. G McKendrick, "Some distributions associated with a randomly arranged set of numbers", *Proc. Roy. Soc. Edinburgh*, Vol 57 (1937), pp. 332-376.

[55] A M Mood, "The distribution theory of runs", *Annals of Math Stat*, Vol. 11 (1940), pp. 367-392.

[56] N. Smirnoff, "Über die Verteilung des allgemeinen Gliedes in der Variationsreihe", *Metron*, Vol. 12, No 2 (1935), pp. 59-81

[57] N Smirnoff, "Sur la dépendance des membres d'une séries de variations", *Bull. Univ. État Moscou, Sér. Int*, *Sect A. Math. et Mécan*, Vol 1, fasc 4, pp. 1-12

[58] W L. Stevens, "Distribution of groups in a sequence of alternatives", *Annals of Eugenics*, Vol. 9 (1939), pp 10-17.

ON THE THEORY OF SAMPLING FROM FINITE POPULATIONS

By Morris H. Hansen and William N. Hurwitz

Bureau of the Census

I—HISTORICAL BASIS FOR MODERN SAMPLING THEORY

The theory for independent random sampling of elements from a population where the unit of sampling and the unit of analysis coincide was developed by Bernoulli more than 200 years ago. The theory that would measure the gains to be had from introducing stratification into sampling was indicated by Poisson a century later. Subsequently, Lexis systematized previous work and provided the theoretical basis for sampling clusters of elements.[1] The adaptation of the work of Bernoulli and Poisson to sampling from finite populations was summarized by Bowley in 1926 [1] approximately a century after the work of Poisson.

An impetus to sampling advancement, following some fundamental statistical contributions of Pearson, Fisher, and others, resulted from the work of Neyman when he published his paper in 1934 on the two different aspects of the representative method [8]. In that paper he introduced new criteria of the optimum use of resources in sampling, including the concept of optimum allocation of sampling units to different strata subject to the restriction that the sample have a fixed total number of sampling units.

If, no matter how a sample be drawn, the cost were dependent entirely on the number of elements included in the sample, there would be little need for theory beyond the classical theories of Bernoulli and Poisson covering the independent random sampling of elements within strata, supplemented by the extension of the theory to finite populations, and the extension to optimum allocation of sampling units. Very often, however, in statistical investigations it is extremely costly, if not impossible, to carry out a plan of independent random sampling of elements in a population. Such sampling, in practice, requires that a listing identifying all the elements of the population be available, and frequently this listing does not exist or is too expensive to get. Even if such a listing is available, the enumeration costs may be excessive if the sample is too widespread. Frequently also, there are other restrictions on the sample design, such as the requirement that enumerators work under the close supervision of a limited number of supervisors, and as a consequence the field operations must be confined to a limited number of administrative centers. Techniques such as cluster sampling [2, 3, 4, 5, 6, 7, 8, 10], subsampling, and double sampling [9], have been

[1] The sampling of clusters of elements refers to the sampling of units that contain more than one element. Examples of cluster sampling include the use of the city block or the county as the sampling unit when the purpose of the survey is to determine the properties of the population made up of individual persons or individual households. In these instances, the city block or county is referred to as the cluster of elements, and the individual person or household is referred to as the element.

developed with the aim of making most effective use of available resources, while keeping within existing administrative restrictions, and thus producing the maximum amount of information possible within these resources and restrictions. Neyman [8], Yates and Zacopanay [10], Cochran [2], Mahalanobis [7], and others have made important contributions in this regard.

We can illustrate a number of the developments indicated above in a simple but fairly general subsampling design. This design involves the sampling of clusters of elements from a stratified population and the subsampling of elements from each of the selected clusters, where the number of elements in each of the primary sampling units within a stratum is the same.

Suppose we have a population made up of L strata, with the i-th stratum containing M_i primary sampling units of N_i elements each. The individual element will be the subsampling unit. Let X_{ijk} be the value of some characteristic of the k-th element of the j-th primary sampling unit in the i-th stratum, and assume that the character to be estimated is

$$\bar{X} = \sum_{i}^{L} \sum_{j}^{M_i} \sum_{k}^{N_i} X_{ijk} \Big/ \sum_{i}^{L} M_i N_i . \tag{1}$$

For example, if $\bar{X}$ is the average income per household in a given city, X_{ijk} might be the income of the k-th household in the j-th city block in the i-th ward; where the household is the subsampling unit, the city block is the primary sampling unit, and the stratification has been by wards. Suppose, further, that we sample m_i primary units from the i-th stratum, and subsample n_i elements from each of the primary units sampled from that stratum.

The "best linear unbiased estimate" [8] of $\bar{X}$ from the sample will be

$$\bar{X}' = \sum_{i}^{L} \frac{M_i N_i}{m_i n_i} \sum_{j}^{m_i} \sum_{k}^{n_i} X_{ijk} \Big/ \sum_{i}^{L} M_i N_i , \tag{2}$$

and the variance of $\bar{X}'$ is

$$\sigma^2_{\bar{X}'} = \sum_{i}^{L} M_i^2 N_i^2 \left\{ \frac{M_i - m_i}{M_i - 1} \frac{\sum_{j}^{M_i} (\bar{X}_{ij} - \bar{X}_i)^2}{M_i m_i} + \frac{N_i - n_i}{N_i - 1} \frac{\sum_{j}^{m_i} \sum_{k}^{N_i} (X_{ijk} - \bar{X}_{ij})^2}{M_i N_i m_i n_i} \right\} \Big/ \left(\sum_{i}^{L} M_i N_i \right)^2 \tag{3}$$

where $\bar{X}_{ij} = \sum_{k}^{N_i} X_{ijk}/N_i$ and $\bar{X}_i = \sum_{j}^{M_i} \sum_{k}^{N_i} X_{ijk}/M_i N_i$.

These formulas have no practical utility in designing samples unless there are, in addition, some considerations of differential costs. Cost relationships sometimes may be stated explicitly as a function of the m_i and the n_i, or, what is frequently the case, they may be approximated sufficiently through intuition and speculation to guide one to a reasonable decision among the various alternatives implied by the design.

If we know the cost function we proceed to determine the values of the m_i and the n_i that make $\sigma^2_{\bar{x}'}$ a minimum for a fixed total expenditure, and also subject to any other restrictions that may be imposed. This theory provides a basis for determining the optimum allocation of the sampling ratios to the various strata, and to primary and secondary sampling units within each stratum.

Such developments, however, must be regarded as only the first step in sample design. We cannot go forward if we only know that the optimum sample design is some particular mathematical function of the population parameters and the cost factors; we need also to know something about the relative magnitudes of certain parameters in the particular populations under consideration, as well as something about the costs associated with the various sampling and estimating operations.

Thus, considerable work in recent years has been done on the study of the relative magnitudes of variances and covariances between and within various types of sampling units and on the study of costs and types of cost functions that operate. Work is being done in this field by the Department of Agriculture in connection with sampling for agricultural items, and is being done also in the Bureau of the Census, and in other places.

II—THE DIRECTION OF MORE RECENT DEVELOPMENTS

The sampling procedure indicated above involves as a first step the definition of the *system of sampling*, such as whether the sampling method will involve cluster sampling, double sampling, or subsampling, and along with this the definition of the stratification and the sampling units. The second step is that of determining the *method of estimation*, together with *the allocation of the sampling units*.

The first step, that of defining the sampling system is taken with a view to administrative feasibility and sampling efficiency, but no simple procedure exists which leads one uniquely to the selection of a system except perhaps by the impractical method of listing and examining all possible alternatives and accepting one on some criterion of best. However, given the definition of a population character to be estimated, and a sampling system, a simple procedure is available that will provide a unique solution to the second step providing we accept some criterion as to what "best" means, such as the best linear unbiased estimate, subject to any cost or administrative restrictions that may be imposed. Such criteria lead us to both our estimating procedure and our allocation of sampling within the sampling system defined.

While no theory with practical applicability has been developed which indicates a "best" system of sampling, and at the same time indicates the "best" estimating procedure and sampling allocation, some progress in the choice of improved sampling systems and estimating procedures has been made. The developments in the following two directions appear to us to be particularly pertinent.

1. Modifications in some of the fairly generally accepted criteria of good

sample estimates have led to more reliable sample results for some types of sampling systems (some of these are mentioned in Sec. III);

2. Some principles are emerging, that have led to improved determination of the sampling units, the strata, and other aspects of the sampling system (some efforts at formulating such principles are reported in Secs. IV, V, and VI).

We shall summarize, principally, some of the recent work in the Census—and in so doing shall mention some work of others that is closely related. Most of the work that we shall summarize relates to problems where the sampling units are clusters of elements and vary in size

III—MODIFICATIONS IN THE CRITERIA FOR GOOD ESTIMATES

The estimate given in the general subsampling problem formulated in Sec. I satisfies the criterion of the "best linear unbiased estimate." Also, as far as our experience has indicated, this estimate is frequently the most efficient one for populations of the form described, that is, where the number of elements in each sampling unit within a stratum is the same. However, if the numbers of elements differ between sampling units, a biased but consistent estimate can frequently be found that has a substantially smaller mean square error[2] than the best linear unbiased estimate.

For example, consider the case where clusters of elements are the sampling units and we want to estimate $\bar{X} = \sum_{i}^{M} X_i / \sum_{i}^{M} N_i$, the average value per element of some specified characteristic. Here M is the number of sampling units in the population, X_i is the aggregate value of the specified character for all elements in the i-th cluster, and N_i is the number of elements in that cluster. The joint distribution of X_i and N_i is unknown, but $\sum_{i}^{M} N_i = N$ is known. Under these circumstances the "best linear unbiased estimate" of $\bar{X}$ from a sample of m clusters turns out to be $\frac{M}{m} \sum_{i}^{m} X_i / N$. However, a smaller mean square error is often obtained by the use of a ratio estimate from the sample such as $\sum_{i}^{m} X_i / \sum_{i}^{m} N_i$. This estimate is excluded by the "best linear unbiased" criterion because it is nonlinear and biased, although the bias is usually negligible and the estimate is consistent. Since the best linear unbiased estimate of $\bar{X}$ requires the knowledge of N, the sample ratio has a further advantage in that it can be used even when N is not known.

A recent paper by Cochran [3] gives a number of consistent though biased esti-

[2] In this paper the terms "mean square error" and "variance" are used interchangeably to refer to $E(X - \hat{X})^2$ when EX is equal to $\hat{X}$, the population character to be estimated. When EX is not equal to $\hat{X}$, however, $E(X - \hat{X})^2$ will be referred to only as the "mean square error" Since, under these latter circumstances, $E(X - \hat{X})^2 = E(X - EX)^2 + (EX - \hat{X})^2$, the mean square error is equal to the variance of X plus the contribution due to the bias.

mates of $\bar{X}$ that make use of the least square estimate of the linear regression of X_i on N_i. These estimates generally have a smaller mean square error than either the best unbiased linear estimate or the simple ratio estimate given above. However, they require knowledge of N, as does the best linear unbiased estimate, and in addition may require detailed tabulations and considerable clerical work as a part of the estimating process.

Both types of biased estimates mentioned above are consistent, and usually have a smaller mean square error than the best linear unbiased estimate for sampling systems in which the sampling units vary in size. Thus, improved sample estimates will be obtained by modifying the "best linear unbiased estimate" criterion to include estimates that are nonlinear, consistent, but have a smaller mean square error than the best linear unbiased estimate.

IV—IMPROVEMENTS IN THE SPECIFICATIONS OF SAMPLING SYSTEMS

A great deal can be done to improve sampling designs through improved specification of the sampling system even though one has only a limited knowledge of the manner in which the population is likely to be made up, and no specific information concerning the particular population parameters involved (see Sec. VI).

1. The sizes of sampling units. A number of recent investigations have indicated the desirability, with costs considered, of keeping the size of cluster very small when clusters of elements are used as the sampling unit in field surveys [2, 5, 6, 7, 8]. It is important to point out, however, that this principle is not necessarily applicable to subsampling systems, and that the use of large clusters as the primary sampling units in a system involving subsampling may yield distinct gains over the use of smaller clusters without subsampling. Moreover, one of the often recurring problems in large-scale studies is the designing of sample surveys within stringent administrative restrictions on the number of different locations in which operations can be carried on. Under such restrictions a procedure commonly used is to choose a limited number of existing political units, such as counties, as the primary sampling units, and then to subsample units such as blocks, small rural areas, or households. Under the circumstances, if the numbers of primary subsampling units to be included in the sample are assumed to be held constant, the use of larger primary sampling units than the existing political units would have the effect of decreasing the sampling variance.

The advantage of using large primary units in subsampling is evident in the simple case when the original units, each having the same number of elements, are consolidated to form half as many enlarged primary units, each twice as large as the original units. The variance between the enlarged primary units will be $\sigma_{2b}^2 = \frac{1}{2}\sigma_{1b}^2(1 + \rho)$, where σ_{1b}^2 is the variance between the original primary units, and ρ is the correlation between the units that are paired. The correlation coeffi-

cient will be close to zero (exactly equal to $-1/\{M-1\}$, where M is the number of original primary units) if the pairing is done at random, and it follows that the variance between counties is then cut at least in half. Ordinarily, ρ will be greater than zero if the paired units are required to be contiguous. However, through choosing for consolidation those contiguous units that are as different as possible, ρ is made as small as possible, and in some instances this minimum value may even be negative. In any event, the smaller the value that ρ takes on, the greater the reduction of the sampling variance between primary units from the use of enlarged units. While the sampling variance within primary units is increased by such consolidations, the increase is slight, and the total sampling variance is almost invariably decreased (see Appendix, Section 1)

The restriction on extending the consolidation of primary units is introduced by the increased cost of subsampling within larger and larger areas This increased cost is to be weighed against the decreased variance If the cost restriction were not sufficiently severe, consolidation would proceed to the point of eliminating the use of primary sampling units altogether, and the subsampling units would be selected independently throughout the entire stratum.

2. Subsampling where the primary units are of unequal size. Use of probability proportionate to size in subsampling. A subsampling system frequently followed, whether or not the primary sampling units vary in size, involves the selection of one or more primary units from each stratum with the probability of selection the same for each primary unit in the stratum, and the subsampling of a fixed proportion of the subsampling units from the selected primary unit. When the primary units vary in size this subsampling system has some administrative disadvantages that arise because the number of subsampling units to be included in the sample will vary with the number of elements in the selected primary unit. (The term "size" of sampling unit as used in this paper refers to the number of elements in the sampling unit.)

The disadvantages in the above system have led in some instances to the specification of a second subsampling system in which, although the primary units were selected with equal probability, the subsampling has been of a constant *number* rather than of a constant proportion.

A third subsampling system that can be recommended over both the above systems is to make the probability of selection of a primary unit proportionate to its size and then to subsample a constant number of subsampling units.

We shall assume that for all three systems only one primary unit is selected from each stratum. Stratification to this degree leads to a smaller sampling variance than does less extensive stratification. For simplicity in making comparisons, we shall assume, furthermore, that the subsampling unit is the element of analysis and that the sample estimate used is of the form $\bar{X}' = \Sigma N_h \bar{X}'_h / \Sigma N_h$ where $\bar{X}'_h$ is the sample average, for the h-th stratum, of the character being estimated, and N_h is the size of that stratum. This estimate, which is frequently used, is biased for the first two systems but unbiased for the recommended sys-

tem. However, an unbiased estimate, say the "best" linear unbiased estimate for the first two systems generally has a much larger mean square error than the biased estimates used in these comparisons and hence has not been considered in the comparisons which follow (see Sec VII, footnote 7).

The first two subsampling systems mentioned are about equally efficient when the number of subsampling units drawn from each primary unit is reasonably large, but each will usually have a larger mean square error than will the recommended system. The difference between the mean square errors of either of the first two and that of the recommended design is given approximately by

$$(4) \qquad \frac{1}{N^2} \sum_h Q_h \bar{N}_h \sigma_h^2 \left[\sum_j \rho_{hj} \bar{N}_h - \sum_j \rho_{hj} N_{hj} \right]$$

where, within the h-th stratum, N_{hj} is the number of elements in the j-th primary sampling unit, $\bar{N}_h$ is the average size of primary sampling unit, Q_h is the number of primary sampling units, ρ_{hj} is the intra-class correlation between elements within the j-th unit and σ_h^2 is the variance between individual elements within the stratum; L is the number of strata (See Section 2 of the Appendix for the development of this difference.)

This difference, which is a multiple of the average covariance between the N_{hj} and ρ_{hj}, will be positive if N_{hj} and ρ_{hj} are negatively correlated, and this is exactly the situation that exists in most practical problems we have encountered in sampling for social and economic statistics (see Sec. VI).

The reduction in the mean square error arises because the recommended design provides a more nearly optimum allocation of sampling as between large and small sampling units than do the other two. It might be possible, of course, as another alternative, to stratify the primary units by size and then allocate sampling to the various strata on the basis of optimum sampling considerations. However, this would mean that some other and perhaps more important modes of stratification would be sacrificed, and moreover, the optimum allocation of sampling between the larger and smaller units could only be guessed at in most practical problems Furthermore, it usually is not possible to stratify on size to the point that there is no variation in the sizes of units within a stratum.

The sample estimate from the recommended system is unbiased whereas the estimates from the other two are usually biased, and sometimes fairly seriously so. (For a proof of this statement see Appendix, Section 1, and see also Sec. VII for a numerical illustration.)

The use of probability proportionate to size serves to decrease only the sampling variation between primary units and has very little effect on the sampling variance within. Therefore, the recommended design shows its greatest advantage over the two alternatives when the contribution of the mean square error between primary units to the total mean square error is large

Ordinarily, the actual sizes of the primary sampling units will not be known, but numbers may be known that are highly correlated with the sizes. For example, ordinarily we will not know the populations of blocks or of cities or

counties at the time a sample is taken, but we may know their populations at the preceding census. Under these circumstances the primary units may be sampled with probabilities proportionate to the previously known (or their estimated) sizes, but if this is done the subsampling is to be modified in order to take account of the changes in the sizes between the two dates. If the actual sizes are known, the constant number taken from the selected primary unit in the h-th stratum is $n_h = t_h N_h$ where t_h is the sampling ratio assigned to the stratum, and N_h is the total number of elements in the stratum. The subsampling ratio within the selected primary unit, therefore, is $t_h N_h / N_{hj}$, where N_{hj} is the number of elements in the selected unit On the other hand, if there is available only a measure of size P_{hj} highly correlated with the actual sizes of the units N_{hj} and, if the probability of selection of the primary unit has been proportionate to the P_{hj} the subsampling ratio in the selected primary unit will be equal to $t_h P_h / P_{hj}$, where P_h is the measure of size of the entire stratum, and P_{hj} is the measure of size of the selected primary unit. The variance of a sample estimate where measures of size are used is given subsequently in this paper (see Eq. (9))

3. The use of area substratification within primary strata in a subsampling system. Another modification, which will be called area substratification within primary strata, may be particularly useful where a relatively small sample is required from a population covering a large area, and where operations must be confined to a limited number of centers.

Some preliminary remarks are necessary before area substratification can be explained. Area substratification requires (a) that the entire population to be sampled be divided into areas that will serve as primary sampling units, (b) that these units be further subdivided into a number of sub-areas; and (c) that certain summary statistical information be available for each of the sub-areas in advance of drawing the sample. The information that must be known for the sub-areas includes a reasonably good measure of their sizes (perhaps the total population, total dwelling units, or total farms) and other information which is indicative of the characteristics of the area, such as whether predominantly farm or nonfarm, predominantly white or colored, etc. The sub-areas, when grouped into homogeneous classes, will serve only to determine the substrata described subsequently, and will not ordinarily serve as the subsampling units, which may be defined independent of the sub-areas.

The definition of the primary sampling units and the classification of them into strata proceed as indicated earlier, with the primary units made as internally heterogeneous as possible within strata that are as homogeneous as possible. It will be assumed that only one primary unit is sampled from each stratum, and that the probability of selecting the j-th primary unit within the h-th stratum is proportionate to P_{hj}, where P_{hj} is the measure of size of the primary unit and is equal to the sum of the measures of size of the sub-areas that it contains. It will be assumed, also, that t_h, the over-all sampling ratio to be used within the h-th stratum, has been determined for all strata on the basis of considerations of optimum allocation.

The introduction of area substratification within primary strata may then be accomplished as follows:

(a) The sub-areas within each primary stratum are classified into substrata on the basis of their characteristics. (For example, they may be classified into predominantly farm and predominantly nonfarm sub-areas, and these further classified on the basis of the average size of farm or average rental value of the dwelling units. In such a case, the sub-areas within the primary stratum that are predominantly farm and that have average rental values lying within a specified interval constitute a substratum.)

(b) The sub-areas within the primary unit selected from each primary stratum are classified into the same substrata.

(c) Subsampling units are defined within each of the substrata within the selected primary units. The number of subsampling units defined within that part of the i-th substratum that is contained within the j-th primary unit is denoted by M_{hij}. (Various types of subsampling units may be defined, such as the individual person, farm, dwelling unit, or structure, a very small area, etc. The subsampling units need be defined only within the selected primary sampling units.)

(d) The number of subsampling units to be included in the sample from the i-th substratum within the selected (j-th) primary sampling unit is

$$m_{hij} = M_{hij} t_h P_{hi} / P_{hij}, \tag{5}$$

where P_{hij} is the measure of size of that part of the i-th substratum that lies within the j-th primary unit, and $P_{hi} = \sum_j P_{hij}$ is the sum of the measures of size of the sub-areas contained in the i-th substratum of the h-th primary stratum. This method of allocating the subsampling provides that the subsample drawn from the selected primary unit is representative, so far as possible, of the entire stratum, rather than of the particular primary unit that happens to be included in the sample from that stratum. To illustrate, suppose the numbers of persons in sub-areas from the 1940 census are used as the measures of their sizes, and that the sub-areas are classified into substrata on the basis of their characteristics in 1940 as indicated by the 1940 Decennial Census of Population. The allocation of the subsampling indicated above then provides that if the proportion of the total population residing in sub-areas that are predominantly farm is 30 percent, the sample will be drawn in such a manner that 30 percent of the 1940 population expected in the sample would be from the predominantly farm sub-areas, even though, in the selected primary sampling unit, perhaps only 15 percent of the 1940 population might reside in such areas.

(e) The population character to be estimated is

$$X = \sum_h^L \sum_i^{S_h} \sum_j^{Q_h} \sum_k^{M_{hij}} X_{hijk}, \tag{6}$$

where X_{hijk} is the aggregate value of a specified characteristic for all of the elements contained within the k-th subsampling unit in the i-th substratum of the j-th primary unit; S_h is the number of substrata and Q_h is the number of primary units in the h-th primary stratum; and L is the number of primary strata. (X might be the total number of workers in the United States, or the total number of farm laborers, etc.) An estimate of X from the sample is

$$X' = \sum_{h}^{L} 1/t_h \sum_{i}^{S_h} \sum_{k}^{m_{hij}} X_{hijk}. \tag{7}$$

No summation over j is involved, because only one primary unit is drawn from the h-th stratum. This is a very simple estimate, involving a sum weighted only at the primary strata level. If the t_h are all set equal to t, i.e , if a constant proportion is sampled from each stratum, the estimate becomes merely the total number of elements in the sample having the specified characteristic multiplied by $1/t$, the reciprocal of the sampling ratio.

The allocation of the subsampling indicated above may be deviated from and the controls of area substratification can still be maintained if proper modifications are made in the sample estimate. In this event, differential weighting must be introduced at the substrata level rather than only for the primary strata.

The definition of heterogeneous primary sampling units, the proper classification of them into strata, and the use of probabilities proportionate to the measures of size in the selection of the primary units are particularly desirable if area substratification is used. If these are not introduced the likelihood of making substantial gains through the use of area substratification is decreased. The definition of the primary strata should be made in conjunction with the definition of the substrata, and should insure that each primary unit has adequate representation of each substratum that is to be defined within that primary stratum. With this restriction observed, the number of significant substrata that can be defined will be limited by the heterogeneity of the primary units Thus, in order to provide for substratification into predominantly farm and predominantly nonfarm areas, the primary sampling units should be defined so that both farm and nonfarm areas are represented in each unit. This procedure not only makes area substratification more effective, but improves the efficiency of the sample in making separate estimates for such classes of the population. However, if this procedure cannot be adhered to exactly in practice, primary units in which certain of the substrata are not represented will occasionally come into the sample. One alternative when this occurs is to combine certain substrata; another is to exclude such primary units from the sample.

Since the number of primary strata is restricted by the number of primary units to be sampled, it is wasteful to set up strata at the primary level with respect to sources of variation that can be controlled adequately through area

substratification. For example, if farm areas and nonfarm areas are to be distinguished in the substrata, the primary strata should not be exhausted by classifying the primary units into a large number of strata by percent farm (percent of total population in primary unit living on farms), since the effect of the substratification is to control the variation in the percentage farm. Limiting the number of percentage farm classes at the primary level makes possible the use of other modes of stratification that will control on farm type, or on the industrial character of the nonfarm population, or on some other similar criteria

Area substratification is to be distinguished from the fairly commonly used method of specifying the number of elements to come into the sample from each of several different classes of elements—whether such quotas are fixed to make the sample correspond with the specified characteristics of the entire primary stratum or of the selected primary sampling unit The method of fixing quotas and instructing interviewers or enumerators to obtain a given number of elements (persons, dwelling units, farms, voters, etc.) having various specified characteristics has a fundamental weakness that is avoided in area substratification within primary strata Such quotas ordinarily must be set on the basis of previous information or rough estimates, and thus cannot accurately reveal changing characteristics of the population Area substratification, on the other hand, uses previous information to insure the proper representation of various types of areas in the sample. The numbers of elements obtained with various specified characteristics are determined from the population as it is, and not as it was at some previous date. In times of rapid change the fixing of quotas on the basis of previous information may introduce increasingly serious biases.

The gain from using previously available information in stratifying areas arises from the fact that there is a high correlation in the characteristic of an area from time to time over a period of several years. An area that is predominantly farm at one date ordinarily will be predominantly farm a few years later. Similarly, while very substantial shifts in population may occur, the numbers of persons in a set of areas at one time ordinarily will be very highly correlated with the numbers a few years later However, area substratification does not depend on the fact that no shifts occur If shifts have occurred it will measure them. If the shifts have been sufficient to completely alter the character of most small areas, it will still provide estimates revealing the changing character of the population, but under these circumstances the efficiency of the method is decreased.

V—EXPECTED VALUES AND VARIANCES FOR THE SUBSAMPLING SYSTEM INCORPORATING THE PRINCIPLES OUTLINED ABOVE

The system of sampling incorporating the principles of enlarged primary units, the selection of primary units with probabilities proportionate to the measures of size and area substratification will be examined more fully below. It will be referred to, for convenience, as the specified subsampling system.

1. The expected value of an estimated total for the specified subsampling system. All summations in the formulas below are over the population unless otherwise indicated. The expected value of X' as defined in Eq. (7) is

$$EX' = \sum_h \sum_j \sum_i \sum_k (1/t_h)(P_{hj}/P_h)(m_{hij}/M_{hij})X_{hijk}.$$

From (5) $t_h = m_{hij}P_{hij}/M_{hij}P_{hi}$, and therefore

$$EX' = \sum_h \sum_j \sum_i \sum_k (P_{hi}/P_{hij})(P_{hj}/P_h)X_{hijk}$$

$$= \sum_h P_h \sum_j \sum_i (P_{hi}/P_h)(P_{hj}/P_h)(X_{hij}/P_{hij}) = \sum P_h R_{h(A)}$$

where

$$P_h = \sum_i P_{hi} = \sum_j P_{hj}; \qquad R_{h(A)} = \sum_j (P_{hj}/P_h)R_{hj(A)};$$

$$R_{hj(A)} = \sum_i (P_{hi}/P_h)R_{hij}; \quad \text{and} \quad R_{hij} = \sum_k X_{hijk}/P_{hij} = X_{hij}/P_{hij}.$$

The $R_{hj(A)}$ will be referred to as the adjusted ratio for the j-th primary unit. It is the weighted average within the j-th unit of the substrata ratios, R_{hij}, where the same set of weights P_{hi} is applied to the R_{hij} in each primary unit within a stratum. The $R_{h(A)}$ is the average, within the h-th stratum of the adjusted ratios. Hence

$$EX' = X + \sum_h P_h(R_{h(A)} - R_h), \tag{8}$$

where

$$R_h = X_h/P_h, \quad \text{with} \quad X_h = \sum_i \sum_j X_{hij},$$

is the ratio of the aggregate value of the specified characteristic for the elements in the h-th stratum to the measure of size of that stratum, and where the population character being estimated (6), is equal to $X = \Sigma X_h = \Sigma P_h R_h$.

From (8), it is seen that X' is a biased estimate of X, although ordinarily, in practice, only slightly so. The bias, equal to $\Sigma P_h(R_{h(A)} - R_h)$, is the sum of the biases for the various primary strata. Under many practical circumstances some of these will be slightly negative and some slightly positive, with the result that the total bias will be relatively small. The bias would be nonexistent if area substratification were not used, or if the form of the sample estimate were properly modified, but here again, as in the case of substituting biased for unbiased estimates discussed in Sec. III, the introduction of a slight bias may result in a substantial reduction in the variance.

A sufficient, although not necessary, condition for the sample estimate (7) with area substratification to be unbiased is for the ratios P_{hij}/P_{hj} to be uncorrelated with the R_{hij} within each substratum. Under these circumstances

$$\sum_j \frac{P_{hj}}{P_h}\frac{P_{hij}}{P_{hj}}R_{hij} = \sum_j \frac{P_{hj}}{P_h}\frac{P_{hij}}{P_{hj}}\sum_j \frac{P_{hj}}{P_h}R_{hij} = \frac{P_{hi}}{P_h}\sum_j \frac{P_{hj}}{P_h}R_{hj}$$

and therefore

$$R_h = \sum_i \sum_j \frac{P_{hj}}{P_h} \frac{P_{hij}}{P_{hj}} R_{hij} = \sum_i \sum_j \frac{P_{hi}}{P_h} \frac{P_{hj}}{P_h} R_{hij} = R_{h(A)} .$$

To illustrate, if the measures of size are the 1940 populations, then the sample estimate will be unbiased if the proportions of the 1940 populations of the primary sampling units that are in the various substrata are uncorrelated with the corresponding R_{hij}. As indicated earlier these conditions are approximated in many practical problems, especially if the primary stratification has been carried out effectively. Moreover, if the conditions are not met approximately, the bias introduced may still be very small. (See Sec. VII for a numerical illustration.)

2. The mean square error of an estimated total for the specified subsampling system. For the development of the mean square error of X' for the specified subsampling system, see the Appendix, Section 2 There it is shown that the mean square error of X' is

$$\sigma^2_{X'} = \sum_h \sum_i \sum_j P_{hi}^2 \frac{P_{hj}}{P_h} \frac{M_{hij} - m_{hij}}{M_{hij} - 1} \frac{\sigma^2_{hij}}{m_{hij} \bar{P}^2_{hij}} + \sum_h P_h^2 \sum_j \frac{P_{hj}}{P_h} (R_{hj(A)} - R_{h(A)})^2 + [\sum P_h (R_{h(A)} - R_h)]^2 \tag{9}$$

where

$$\sigma^2_{hij} = \sum_k (X_{hijk} - \bar{X}_{hij})^2 / M_{hij}$$

is the variance between subsampling units within a substratum of the aggregate value of a specified characteristic for the subsampling unit and

$$\bar{P}_{hij} = P_{hij} / M_{hij}$$

is the average measure of size of the subsampling units in the h-i-j-th area.

The first term of (9) is the contribution of the variance between subsampling units and may be kept small by proper definition of the subsampling units, and, of course, by increasing the subsampling ratio. The second term of (9) is the contribution of the variance between primary sampling units within strata; and the third term is the contribution of the bias, which, as indicated before, ordinarily will be of negligible size, so that the mean square error and the variance will be approximately equal.

It is the variance between primary sampling units that contributes most heavily to the total variance in many subsampling situations, and it is on this contribution that the modifications proposed in this paper have their principal effect. The effect of area substratification is seen by comparing the variance between primary units given above with that obtained if area substratification were not used but other aspects of the design remained unchanged. In this

event the variance between primary units involves the variance of the ratio, $R_{hj} = \sum_i X_{hij}/P_{hj} = X_{hj}/P_{hj}$, instead of the variance of the adjusted ratio, $R_{hj(A)}$.

The relationship between the variance of R_{hj} and that of $R_{hj(A)}$ within the h-th primary stratum is given by

$$\sigma^2_{R_{hj}} = \sigma^2_{R_{hj(A)}} + \sigma^2_{R_{hj}-R_{hj(A)}} + 2\rho\,\sigma_{R_{hj(A)}}\,\sigma_{R_{hj}-R_{hj(A)}}, \tag{10}$$

where $\sigma^2_{R_{hj}-R_{hj(A)}}$ is the variance of the difference between the adjusted and the unadjusted ratios, and ρ is the correlation between the adjusted ratio and the amount of the adjustment. Thus, if the correlation is near zero or positive, there will be a gain from the introduction of area substratification, although there may be a loss if the correlation is highly negative. Essentially, the condition for ρ being equal to or near zero is the same as that for the sample estimate being unbiased, namely, that the P_{hij}/P_{hj} be uncorrelated or only slightly correlated with the R_{hij} within each substratum.[3]

The variance of $R_{hj(A)}$ rather than that of R_{hj} occurs in the variance of X' because the subsampling numbers were allocated proportionate to the P_{hi}, no matter what primary sampling unit happened to be selected for inclusion in the sample. The ratio R_{hj}, like $R_{hj(A)}$ may be regarded as the weighted average of the R_{hij} but with the weights equal to P_{hij} instead of P_{hi}, and thus varying from primary unit to primary unit. It would appear, therefore, from the relationship of the variances given above, that if the substrata are effective, and if the P_{hij} are highly correlated with the actual sizes of the substrata, the weighted average using fixed weights in all primary units should have a considerably smaller variance than that using variable weights. This turns out to be the case in many practical situations, some illustrations of which will be given later (see Sec. VII).

3. The mean square error of ratio estimates for the specified subsampling system. The need for estimating a ratio from a sample arises in two cases, first, when the ratio is the population character for which an estimate is desired, and second, when the application of a ratio from the sample to a known total uses additional available information for obtaining an improved estimate of the desired total.

Ratio estimates are desired as an end-result when, for example, the change in a characteristic from one time to another is being considered. Thus, if Y' is the estimated total income of farm workers at one date, and X' the corresponding estimated total income at a second date, then $r' = X'/Y'$ is an estimate of the relative change in the total income of farm workers over the period of time covered. Similarly, the estimate of a percentage such as the percentage of the

[3] Actually, a sufficient, although not necessary, condition for ρ to be equal to zero is that P_{hij}/P_{hj} be uncorrelated with both the ratio R_{hij} and the cross-product $R_{hij}\,R_{hgj}$ for all pairs of substrata

workers unemployed will involve the ratio of two random variables from the sample Ratio estimates from a sample may be particularly useful in instances where the reliability of the ratio estimate is greater than the reliability of the estimate of either the numerator or the denominator, as is frequently the case.

Ratio estimates may be used as a means of obtaining an estimated aggregate value of a specified characteristic, if Y, the aggregate value of a second characteristic highly correlated with X is known exactly from independent sources, and X' and Y', estimates of X and Y respectively, are available from the sample. Thus

$$X'' = [X'/Y']Y = r'Y \tag{11}$$

is an estimate of the aggregate value of the specified characteristic. If the correlation, in successive samples, between X' and Y' is sufficiently high, the ratio estimate will be a more efficient estimate of X than will X', the simple estimated total given earlier (7), but X' will prove the more reliable estimate when the correlation is low.[4] Thus, X'', when the correlation between X' and Y' is sufficiently high, makes use of more of the relevant available information for estimating X than does X'.

The application of ratio estimates to the specified subsampling system is considered below.

(*a*) *The estimated ratio and its mean square error.* The estimate of the population ratio $r = X/Y$ is:

$$r' = \frac{X'}{Y'} = \frac{\sum_h^L \frac{1}{t_h} \sum_i^{S_h} \sum_j^1 \sum_k^{m_{hij}} X_{hijk}}{\sum_h^L \frac{1}{t_h} \sum_i^{S_h} \sum_j^1 \sum_k^{m_{hij}} Y_{hijk}}, \tag{12}$$

where X' is given in (7) above, and Y' is a similar estimate of the total value of a second characteristic. The mean square error of r' is approximately

$$\begin{aligned} \sigma_{r'}^2 = \frac{1}{\overline{Y}^2} \Bigg\{ & \sum_h \sum_i \sum_j P_{hi}^2 \frac{P_{hj}}{P_h} \frac{M_{hij} - m_{hij}}{M_{hij} - 1} \frac{\sum_k Y_{hijk}^2 (r_{hijk} - r_{hij})^2}{m_{hij} M_{hij} \overline{P}_{hij}^2} \\ & + \sum_h \sum_i \sum_j P_{hi}^2 \frac{P_{hj}}{P_h} \frac{M_{hij} - m_{hij}}{M_{hij} - 1} \sigma_{hij \cdot Y}^2 \frac{(r_{hij} - r)^2}{m_{hij} \overline{P}_{hij}^2} \\ & + \sum_h P_h^2 \sum_j \frac{P_{hj}}{P_h} R_{hj(A)\,Y}^2 (\bar{r}_{hj(A)} - \bar{r}_{h(A)})^2 \\ & + \sum_h P_h^2 (\bar{r}_{h(A)} - r)^2 \sum_j \frac{P_{hj}}{P_h} (R_{hj(A):Y} - R_{h(A)\,Y})^2 \Bigg\} \end{aligned} \tag{13}$$

[4] The variance of the ratio of random variables of the form $r' = X'/Y'$ is approximately $\sigma_{r'}^2 = r^2(V_{X'}^2 + V_{Y'}^2 - 2\rho_{X'Y'}V_{X'}V_{Y'})$ where V indicates the coefficient of variation of the variable designated by the subscript, and $\rho_{X'Y'}$ is the correlation Hence, if $\rho_{X'Y'}$ is sufficiently large $V_{r'}^2$ will be less than $V_{X'}^2$. The size of $\rho_{X'Y'}$ required depends on the relative magnitudes of the coefficients of variation of X' and Y'.

where

X_{hijk} = the aggregate value of a specified characteristic for the elements in the k-th subsampling unit within the h-i-j-th area, for which a total is to be estimated;

Y_{hijk} = the aggregate value of a second specified characteristic for the elements in the same subsampling unit, and for which the total in the population is known;

$$Y_{hij} = \sum_k Y_{hijk}, \quad \text{and} \quad Y_h = \sum_i \sum_j Y_{hij}.$$

$$\sigma^2_{hij;Y} = \frac{\sum_k (Y_{hijk} - \bar{Y}_{hij})^2}{M_{hij}}$$ is the variance of the sampling units in the h-i-j-th area with respect to the second characteristic, and $\bar{Y}_{hij} = Y_{hij}/M_{hij}$.

$$R_{hj(A);Y} = \sum \frac{P_{hi}}{P_h} \frac{Y_{hij}}{P_{hij}}$$ is the adjusted average of the Y_{hij}, and

$$r_{hijk} = \frac{X_{hijk}}{Y_{hijk}}, \quad r_{hij} = \frac{X_{hij}}{Y_{hij}},$$ etc., are the ratios of the X to the Y for the areas indicated by the subscripts, and

$$\bar{r}_{hj(A)} = \frac{R_{hj(A)}}{R_{hj(A);Y}}, \quad \text{and} \quad \bar{r}_{h(A)} = \frac{R_{h(A)}}{R_{h(A)\ Y}}$$ are the ratios of the adjusted ratios for X and Y indicated by the subscripts;

and the remaining symbols are as defined in the sections above where the expected value and variance of X' are given.

The first and third terms of (13) are, ordinarily, the principal contributing terms The second and fourth terms contain contributions due to the variation between the means of the substrata and the primary strata respectively even though the sample was stratified with respect to these classes. In some instances, the contributions of these terms will be important. The between strata contributions arise because the primary and subsampling units vary in size with respect to the character Y.

This formula for the mean square error of a ratio is approximately equal to the one more commonly used given in footnote 4. The two formulas, both of which are approximations, would be identical if certain terms which are ordinarily negligible were retained in (13). This latter formula has the advantage of indicating the effect of different aspects of the design of the sample on the variance of the ratio. The derivation of this approximate variance formula is given in the Appendix, Section 3, together with an indication of the accuracy of the approximation.

(b) *The estimated totals and their mean square errors.* As mentioned earlier, two estimates of X, the aggregate value of a given characteristic for all elements are X' (7), and X'' (11) The mean square error of X' is given by (9) and that of X'' is simply equal to $Y^2\sigma^2_{r'}$, where $\sigma^2_{r'}$ is given approximately by (13).

The decision as to whether to use X' or X'' as an estimate of X depends, of course, in the first instance, on whether Y is known, and in the second instance, on the relative magnitudes of the respective mean square errors given in (9) and (13). These may be approximated from prior knowledge concerning the relationships in the population under investigation, or they may be estimated from preliminary sample investigations. However, in instances where there is a positive correlation between the $X_{hij k}$ and the $Y_{hij k}$ within substrata, it is fairly safe to assume that if the information necessary for the ratio estimate is available, there will be little to lose and possibly considerable to gain from its use.

The use of (11) instead of (7) is often desirable when Y in (11) is the aggregate value of the actual sizes of the primary units, and Y' is its estimate. This is particularly so if the measures of size used are not fairly precise measures of the actual sizes, and if, at the same time, the actual size is highly correlated with the character being estimated, in which case the use of ratio estimates will yield gains in both the between primary unit contribution and the within primary unit variance. (See Sec. VII for numerical illustrations.) However, if the measures of size are identical with the actual sizes (i.e., $P_{hij k} = Y_{hij k}$) the last two terms of (13) are identical with the between primary unit contribution to the variance of X' (9), and only the within primary unit variance is affected by the ratio estimate.

While it is fairly safe in practice, if Y is known, to make use of X'' instead of X' as the estimate of X, some care must be exercised to make sure that the $X_{hij k}$ has at least a moderately high average correlation with the $Y_{hij k}$, where the correlations considered are those within substrata within primary sampling units. If this correlation is low, and if the size of the subsampling unit varies considerably, the ratio estimate may be considerably less efficient than the simple total estimate. On the other hand, if the measures of size of the various substrata and of the primary sampling units are fairly close measures of the actual size, and if the subsampling units have been carefully defined so that they do not vary too greatly in size, the two estimates are likely to have about the same efficiency.

VI—SOME PHYSICAL PROPERTIES OF FREQUENTLY OCCURRING POPULATIONS THAT ARE BASIC TO THE SAMPLING PRINCIPLES RECOMMENDED IN THIS PAPER

Many actual populations are characterized by the following physical properties:

(i) The elements within a cluster are positively correlated with regard to a specified characteristic.

(ii) Clusters containing large numbers of elements have greater internal heterogeneity than clusters containing small numbers of elements.

(iii) Increasing the size of the cluster brings in correlated elements (e.g., in population or agriculture surveys larger clusters are formed by including households or farms in adjacent areas)

The first of these properties is recognized implicitly in the literature where the losses of efficiency through the use of large clusters as sampling units are frequently cited. In our experience the second and third properties hold just as commonly in actual populations, and ordinarily for the same populations for which the first property holds.

The presence of these physical properties in combination within strata leads to the following mathematical relationships that have been used throughout this paper:

(a) The sizes of the primary sampling units, N_{hj}, are negatively correlated with the ρ_{hj}, the intra-class correlations between elements within the units;

(b) The N_{hj} and $N_{hj}\rho_{hj}$ are positively correlated,

(c) The N_{hj} and σ^2_{hj} are positively correlated;

(d) The N_{hj} and σ^2_{hj}/N_{hj} are negatively correlated.

The use of these relationships has determined most of the choices among alternative procedures throughout this paper. The relationships, of course, do not necessarily hold, and exceptions to them can be found [5]. The frequent occurrence of populations characterized by such properties justifies further research on the more effective use of these and other properties that may be found to hold.

VII--SOME APPLICATIONS OF THE PRINCIPLES DESCRIBED IN THIS PAPER TO AN ACTUAL SAMPLING PROBLEM

The analyses summarized below were carried out for the purpose of deciding between alternative sampling procedures in the revision of a monthly national sample for labor force and other characteristics Budgetary and administrative restrictions made it necessary to confine the field operations to a limited number of administrative centers scattered over the country, from which a sample of less than one-tenth of one percent of the population of the United States was to be drawn.

The original sample (the one to be revised) was of a usual subsampling design in which counties were used as the primary sampling units, and households or small clusters of households were used as the subsampling units. In the revised sample contiguous counties were combined wherever administratively feasible, to form more heterogeneous primary units than the individual counties. Approximately 2000 primary sampling units were formed from the 3000 counties in the United States. The combinations of counties, the primary stratification, the area substratification, and the measures of size, were determined on the basis of 1940 Decennial Census data together with more recent data where available.[5]

The applications of the various principles suggested in this paper have been

[5] See [11] for a full description of the proposed revised sample, including an outline of the criteria of stratification used. That paper may be useful as a simple description of an application of the specified subsampling system.

evaluated by estimating 1930 Census labor force characteristics from a sample that was stratified on the basis of 1940 and more recent data. This constituted a particularly severe test of some of the methods, because of the substantial shifts that had taken place during the 10-year interval between 1930 and 1940

The analyses to be summarized in this section are concerned primarily with the gains obtainable under favorable circumstances by the introduction of three sampling principles, namely,

(1) enlarged primary units (see Sec. IV-1);
(2) the sampling of primary units with probability proportionate to measures of their size (see Sec. IV-2);
(3) area substratification (see Sec. IV-3).

Some comparisons are also given to illustrate the effect of using alternative sample estimating formulas. Computations have been made for six of the principal items that are currently being included in a monthly report of the labor force; namely, total numbers of male and female workers, total numbers of male and female agricultural workers, and total numbers of male and female nonagricultural workers The comparisons between alternative systems have been made holding constant both the primary stratification criteria and the expected numbers of persons to be drawn into the sample.

The percentage gains given below are the reductions in the *between* primary unit contributions (which include the bias contributions) to the mean square error.[6] Except where otherwise specified, the sample estimate used is given by (7).

1. Gains obtained by introducing enlarged primary units. The gains obtained by using enlarged primary units are calculated by comparing the mean square errors arising from the sampling design in which individual counties are primary units with the mean square errors arising from the design in which combinations of counties are the primary units. In both designs, the primary units are drawn with equal probabilities and no area substratification is used. For this comparison, preliminary computations have been completed for only a limited number of strata and for two of the labor force items given above; namely, total male workers and total female workers. The reduction in the sampling errors obtained by introducing enlarged primary units is estimated to be 48 per cent for total male workers and 26 per cent for total female workers.

2. Further gains obtained by introducing probability proportionate to measures of size. The further gains obtained by using the principle of sampling with probability proportionate to measures of size are calculated by comparing the mean square errors arising from the design in which the units are drawn with

[6] The contribution of the variance *within* the primary units to the total mean square error was relatively small in all instances, and practically unaffected by the introduction of the various principles.

equal probability with the mean square errors arising from the design in which the units are drawn with probability proportionate to measures of size. In both the designs, the primary units are combinations of counties, and in neither of them is area substratification used. The estimated per cent gains are as follows:

Total Workers		*Agricultural Workers*		*Nonagricultural Workers*	
Male	*Female*	*Male*	*Female*	*Male*	*Female*
50	8	77	6	19	21

The gains reflect both decreases in the sampling variance and the elimination of the bias which arises when the primary units are drawn with equal probabilities.[7]

3. Further gains obtained by introducing area substratification. The further gains obtained by using the principle of area substratification are calculated by comparing the mean square errors for the design in which area substratification is not used, with those for the design in which area substratification is introduced. In both these designs the primary units are combinations of counties, and are drawn with probability of selection proportionate to measures of their sizes. The estimated per cent gains are as follows:

Total Workers		*Agricultural Workers*		*Nonagricultural Workers*	
Male	*Female*	*Male*	*Female*	*Male*	*Female*
6	31	46	51	32	22

4. Gains obtained by the integration of the above principles into a single subsampling system (the specified subsampling system). The gains obtained by using all three principles are calculated by comparing the mean square errors for the specified subsampling system (in which all three principles are used) with the mean square errors for the system in which none of these principles is used. In the specified subsampling system, combined counties are the primary units, the primary units are drawn with probability proportionate to measures of their size, and area substratification is used. In the other system, the primary units are individual counties, the sampling is done with equal probabilities and area substratification is not used. Preliminary computations for this comparison are available for only 2 of the 6 labor force items; namely, total male and total female workers. The estimated gains were 76 per cent for male workers and 53 per cent for female workers.

[7] As indicated before, estimate (7) is used in both designs compared above. This estimate is unbiased for the design in which the primary units are drawn with probability proportionate to measures of size, but is biased for the design in which they are drawn with equal probabilities. However, for the latter design, the biased estimate is usually much more efficient than the best linear unbiased estimate. For the six labor force items, the best linear unbiased estimate gives rise to variances that are several times as large as the mean square errors for the biased estimate.

Calculations are available for all 6 items to measure the gains obtained by the use of the last two of the principles in combination; namely, probability proportionate to measures of size and area substratification. For measuring these gains, the systems are as described above, except that in both designs the primary units are combinations of counties. The estimated per cent gains are as follows:

Total Workers		*Agricultural Workers*		*Nonagricultural Workers*	
Male	*Female*	*Male*	*Female*	*Male*	*Female*
54	37	88	54	45	39

While both the specified subsampling system and the alternative to which it was just compared are biased designs, the bias in the specified system is appreciably smaller than the bias in the latter. For example, while the bias of the specified system in the estimation of total male workers was less than one-half per cent of the true total male workers, the bias for the alternative design in the estimation of the same population character was more than one and one-half per cent.

5. The choice of estimate to use with the specified subsampling system. The simple estimate (7) given for the specified subsampling system may be improved on by the use of regression techniques (see Sec. III). However, such techniques may require a great deal of clerical work, so that they frequently cannot be used in practice. As indicated in the last part of Sec. V, however, if certain independent information such as a knowledge of the total population is available, a simple ratio estimate of the form of (12) may sometimes introduce gains over (7). The use of the ratio estimate may be particularly desirable when the correlation between the measures of size and the actual sizes of the primary sampling units is only moderately high, and when, at the same time, the actual sizes are highly correlated with the values for the character being estimated. A small-scale experiment in the sampling for labor force items indicated that for estimating total male workers for 1930, both the variance between primary units and the variance within primary units for the ratio estimate (12) were approximately one-half that for the simple estimate (7). The use of the ratio estimate had very little effect in the estimation of the remaining five labor force characteristics. The reduction in variance of the total male employment figure was brought about because migration since 1930 reduced the correlation between the 1930 and 1940 sizes, and furthermore, the number of male workers is highly correlated with the total population. Similar reductions for the variances of the other five items were not obtained because the correlations with actual sizes for the other items were not as high.

6. Some final remarks. The gains just obtained arose from application of the sampling principles enumerated above. The situations that these principles were applied to are favorable, but are frequently met in practice. The principles differ in their effect depending on the particular attributes of the population

being studied. The use of enlarged primary units may be desirable whenever the enlarged units are internally more heterogeneous than are the smaller units. The selection of primary units with probability proportionate to size is desirable for the general classes of populations described in Sec. VI whenever the primary units vary considerably in size The use of area substratification is limited to sampling situations where large primary units are used. The joint effect of all three principles shows to greatest advantage when subsampling is used, the primary units are large, but variable in size, and the number of primary units included in the sample is limited by cost or administrative conditions. The types of estimates described in Sec. III may be effective in a large number of physical situations other than those mentioned in this paper.

ACKNOWLEDGMENT

The authors are particularly indebted to Harold Nisselson for many helpful suggestions, and also for planning and directing the analyses summarized in Sec. VII. They are heavily indebted also, to W. G. Cochran, W. Edwards Deming, L. R. Frankel, William G. Madow and Frederick F. Stephan for reading the manuscript and making many helpful suggestions. The work has been carried out in the Bureau of the Census, in connection with the development of various sampling surveys, under the general direction of J. C. Capt, Director, and Philip M. Hauser, Assistant Director.

APPENDIX

1. The effect of the consolidation of the primary units on the sampling variance (see Sec. IV-1). Let $\bar{X}_1' = \sum_{j}^{q} \sum_{k}^{n} X_{jk}/qn$, be the average for the sample where the primary units are the original units and where X_{jk} is the value of the k-th element in the j-th primary unit; q is the number of primary units in the sample, and n is the number of elements sampled from each of the q primary units. The variance of $\bar{X}_1'$ is

$$\sigma^2_{\bar{x}_1} = \frac{N-n}{(N-1)nq}\,\sigma^2_{1w} + \frac{Q-q}{(Q-1)q}\,\sigma^2_{1b} \tag{14}$$

where Q is the number of original primary units in the population; N is the number of elements in each original primary unit; $\sigma^2_{1w} = \Sigma\Sigma(X_{jk} - \bar{X}_j)^2/QN$ is the variance within the original primary units, with $\bar{X}_j = \sum_{k} X_{jk}/N$; and $\sigma^2_{1b} = \Sigma(\bar{X}_j - \bar{X})^2/Q$ is the variance between the original primary units, with $\bar{X} = \Sigma\bar{X}_j/Q$.

$$\sigma^2 = \Sigma\Sigma(X_{jk} - \bar{X})^2/QN = \sigma^2_{1w} + \sigma^2_{1b}. \text{ Then} \tag{15}$$

$$\sigma^2_{1b} = \sigma^2[1 + \rho_1(N-1)/N \tag{16}$$

where $\rho_1 = \left[\sigma_{1b}^2 - \frac{\sigma_{1w}^2}{N-1}\right]\frac{1}{\sigma^2}$ is the intra-class correlation[8] between elements in the original units.
From (15) and (16)

$$\sigma_{1w}^2 = \frac{N-1}{N}\sigma^2(1-\rho_1). \tag{17}$$

Hence

$$\sigma^2_{\bar{x}'} = \frac{N-n}{N}\frac{\sigma^2}{nq}(1-\rho_1) + \frac{Q-q}{(Q-1)q}\frac{\sigma^2}{N}[1+\rho_1(N-1)]. \tag{18}$$

Similarly, the variance of $\bar{X}_2'$ is

$$\sigma^2_{\bar{X}_2'} = \frac{CN-n}{CN}\frac{\sigma^2}{nq}(1-\rho_2) + \frac{Q-qC}{(Q-C)q}\frac{\sigma^2}{CN}[1+\rho_2(CN-1)] \tag{19}$$

where $\bar{X}_2'$ is the mean for the enlarged primary units, ρ_2 is the intra-class correlation between elements in the enlarged primary units and C is the number of original units combined to form each enlarged unit. Then

$$\sigma^2_{\bar{X}_1'} - \sigma^2_{\bar{X}_2} = \frac{\sigma^2}{qN}\left\{\frac{(q-1)(C-1)}{(Q-1)(Q-C)} + \rho_1 a_1 - \rho_2 a_2\right\} \tag{20}$$

where $a_1 = \frac{(Q-q)(N-1)}{Q-1} - \frac{N-n}{n}$ and $a_2 = \frac{(Q-Cq)(CN-1)}{(Q-C)C} - \frac{CN-n}{Cn}$.

Since

$$a_1 - a_2 = \frac{(C-1)(q-1)(QN-1)}{(Q-1)(Q-C)} \geqq 0 \quad \text{and} \quad \frac{(q-1)(C-1)}{(Q-1)(Q-C)} \geqq 0,$$

then a gain is brought about by enlarging primary units whenever $\rho_1 > \rho_2$, where ρ_1 and ρ_2 are both positive

2. Comparison of variances of certain alternative subsampling systems where the primary units are of unequal sizes. The development of (4), the formula for the difference between the variances of sample estimates compared in Sec. IV-2 is given below. We shall confine ourselves to the simple case where only one primary sampling unit is drawn into the sample from each stratum Let

$$\bar{X}' = \Sigma N_h \bar{X}_h'/N \tag{21}$$

be the sample estimate used for each of the three designs to be compared, where $\bar{X}_h' = \bar{X}_{hj}' = \sum_k^{n_{hj}} X_{hjk}/n_{hj}$, and X_{hjk} is the value of the k-th element in the j-th

[8] For definitions and properties of intra-class correlations, see Secs. 38-40 of *Statistical Methods for Research Workers*, R. A Fisher, and [5].

primary unit in the h-th stratum; L is the number of strata; n_{hj} is the number of elements drawn into the sample from the j-th primary unit in the h-th stratum with N_{hj} the corresponding total number, $N_h = \sum_j^{Q_h} N_{hj}$ with Q_h = the number of primary units in the h-th stratum, and $N = \sum_h^L N_h$. If the subsampling within a stratum is of a constant proportion, C, as in the first of the subsampling systems mentioned, n_{hj} in the above estimate is equal to $C\,N_{hj}$. If the subsampling within a stratum is of a constant number, as in the second subsampling system mentioned, as well as in the recommended system, n_h, is equal to $\bar{n}_h = C\sum_j N_{hj}/Q_h = C\bar{N}_h$.

We shall denote the sample estimate for the first design by $\bar{x}_1'$, that for the second design by $\bar{x}_2'$, and that for the recommended design by $\bar{x}_3'$.

The expected values of the sample estimates for the first two designs, $\bar{x}_1'$, and $\bar{x}_2'$, are the same, and are equal to

$$E\bar{x}_1' = E\bar{x}_2' = \bar{X} = \frac{1}{N}\sum_h \frac{N_h}{Q_h}\sum_j \frac{n_{hj}}{N_{hj}}\sum_k \frac{X_{hkj}}{n_{hj}} = \frac{1}{N}\sum_h \frac{N_h}{Q_h}\sum_j \bar{X}_{hj}$$

where $\bar{X}_{hj} = \sum_k X_{hjk}/N_{hj}$. Thus, since $\bar{X}$ is not, in general, equal to $\sum_{h,s,j} X_{hsj}/\sum_{h,j} N_{hj} = \bar{X}$, both $\bar{x}_1'$ and $\bar{x}_2'$ are biased estimates of $\bar{X}$.

For the recommended design, in which the primary unit is drawn with probability of selection proportionate to size and a constant number taken from the sampled units within a stratum, the expected value of the sample estimate is

$$E\bar{x}_3' = \frac{1}{N}\sum_h\sum_j N_h \frac{N_{hj}}{N_h}\frac{\bar{n}_h}{N_{hj}}\frac{X_{hj}}{\bar{n}_h} = \frac{1}{N}\sum_h\sum_j X_{hj} = \bar{X} \tag{22}$$

and therefore the estimate for the recommended design is unbiased.

The mean square error of $\bar{x}_1'$ is

$$\sigma^2_{\bar{x}_1'} = \frac{1}{N^2}\sum_h \frac{N_h^2}{Q_h}\left[\sum_j \frac{N_{hj}-n_{hj}}{(N_{hj}-1)n_{hj}}\sigma_{hj}^2 + \sum_j (\bar{X}_{hj}-\bar{X}_h)^2\right] + (\bar{\bar{X}}-\bar{X})^2 - \frac{1}{N^2}\sum_h N_h^2(\bar{\bar{X}}_h - \bar{X}_h)^2 \tag{23}$$

where $\sigma_{hj}^2 = \sum_k (X_{hjk}-\bar{X}_{hj})^2/N_{hj}$ is the variance between elements within the j-th primary sampling unit of the h-th stratum, $\bar{\bar{X}}_h = \sum_j \bar{X}_{hj}/Q_h$, $\bar{X}_{hj} = X_{hj}/N_{hj}$, and $\bar{X}_h = \sum_j\sum_k X_{hjk}/\sum_j N_{hj} = \sum_j N_{hj}\bar{X}_{hj}/\sum_j N_{hj}$. The first term in the square bracket of (23) is the contribution of the variance within primary units. The second term in the square bracket is an approximation to the mean square error between primary units and the remaining terms give the error in this approximation. The mean square error of $\bar{x}_2'$ is given by the same formula but with n_h, replaced by $\bar{n}_h$.

The difference between $\sigma^2_{\bar{x}'_1}$ and $\sigma^2_{\bar{x}'_2}$ is

$$\sigma^2_{\bar{x}'_1} - \sigma^2_{\bar{x}'_2} = \frac{1}{CN^2} \sum_h \frac{N_h^2}{Q_h} \sum_j \sigma^2_{hj} \frac{N_{hj}}{N_{hj} - 1} \left(\frac{1}{N_{hj}} - \frac{1}{\bar{N}_h} \right), \tag{24}$$

which will be positive if σ^2_{hj}/N_{hj} is negatively correlated with N_{hj}, as is almost invariably the case in practice (see Sec. VI). Thus, since $\sigma^2_{\bar{x}'_1}$ ordinarily is larger than $\sigma^2_{\bar{x}'_2}$, it will suffice to compare $\sigma^2_{\bar{x}'_2}$ with $\sigma^2_{\bar{x}'_3}$ to show that the recommended subsampling system is more efficient than either of the first two mentioned.

The variance for the recommended design is

$$\sigma^2_{\bar{x}'_3} = \frac{1}{N^2} \sum_h N_h^2 \left[\sum_j \frac{N_{hj}}{N_h} \frac{N_{hj} - \bar{n}_h}{N_{hj} - 1} \frac{\sigma^2_{hj}}{\bar{n}_h} + \sum_j \frac{N_{hj}}{N_h} (\bar{X}_{hj} - \bar{X}_h)^2 \right]. \tag{25}$$

For comparing the mean square error of $\bar{X}'_2$ with the variance of $\bar{X}'_3$ we shall define

$$\rho_{hj} = \frac{1}{\sigma^2_h} \left[(\bar{X}_{hj} - \bar{X}_h)^2 - \frac{\sigma^2_{hj}}{N_{hj} - 1} \right]$$

as the intra-class correlation coefficient between elements within the j-th primary unit, where σ^2_h is the variance between all elements within the h-th stratum. In this comparison, the terms outside the square brackets in (23), have been ignored because their contribution to the mean square error is either positive or negligible. Then,

$$\sigma^2_{\bar{x}'_2} - \sigma^2_{\bar{x}'_3} = \frac{1}{N^2} \sum_h \frac{N_h^2}{Q_h} \left\{ \sum_j \frac{N_{hj}}{N_{hj} - 1} \frac{\sigma^2_{hj}}{\bar{n}_h} \left(1 - \frac{N_{hj}}{\bar{N}_h} \right) + \sigma^2_h \sum_j \rho_{hj} \left(1 - \frac{N_{hj}}{\bar{N}_h} \right) \right\}. \tag{26}$$

The second term of this difference was given in Sec. IV-2 as the approximate difference, and the first term was neglected. To examine the relative magnitudes of the two terms we shall write

$$\frac{N_{hj}}{N_{hj} - 1} \sigma^2_{hj} = \sigma^2_h (1 - \delta_{hj}). \tag{27}$$

Then

$$\sigma^2_{\bar{x}'_2} - \sigma^2_{\bar{x}'_3} = \frac{1}{N^2} \sum_h \frac{N_h^2}{Q_h} \sigma^2_h \left\{ \frac{1}{\bar{n}_h} \sum_j \delta_{hj} \left(\frac{N_{hj}}{\bar{N}_h} - 1 \right) - \sum_j \rho_{hj} \left(\frac{N_{hj}}{\bar{N}_h} - 1 \right) \right\}. \tag{28}$$

For the general class of populations given in Sec. VI the covariance between δ_{hj} and N_{hj}, and also that between ρ_{hj} and N_{hj}, will be negative. Moreover, in many practical problems of this class the two covariances will be of approximately the same magnitude. In such instances the first term of (27) will be equal to $\frac{1}{\bar{n}_h}$ times the second, and thus smaller than the second term for all $\bar{n}_h > 1$, and much smaller for moderately large values of $\bar{n}_h$. For example, in popula-

tions made up of clusters of different sizes for which the conditional probability of an element having a particular property for a fixed size of cluster is the same for all sizes of clusters, the two covariances will be very nearly equal. A number of practical problems approximate this situation. Moreover, even in the situations where the covariance of δ_{hj} and N_{hj} is several times that of ρ_{hj} and N_{hj}, say 5 times as large, then the second term will be larger than the first for all $\bar{n}_h > 5$.

Some numerical illustrations of the gains obtained through the use of the recommended system are given in Sec. VII, and for some of the items for which results are summarized in that section the gains were substantial.

3. The derivation of the variance formulas (13) and (9). The mean square error of a ratio of random variables is generally approximated from Taylor's expansion If X' and Y' are random variables, $Y' > 0$, and r is the population character of which $X'/Y' = r'$ is an estimate, then

$$(29) \qquad E\left\{\frac{X'}{Y'} - r\right\}^2 = E\,\frac{Y'^2}{(EY')^2}\left(\frac{X'}{Y'} - r\right)^2 + E\left(1 - \frac{Y'^2}{(EY')^2}\right)\left(\frac{X'}{Y'} - r\right)^2.$$

The first term in the right-hand side of (29) is a first approximation to the mean square error from Taylor's expansion, and the second term is the error in this approximation.

Eq. (13), and as a special case (9), is derived as follows:

$$(30) \qquad E(r' - r)^2 = E\left\{\frac{\sum_h^L \frac{1}{t_h}\sum_i^{S_h}\sum_j^{1}\sum_k^{m_{hij}} X_{hijk}}{\sum_h^L \frac{1}{t_h}\sum_i^{S_h}\sum_j^{1}\sum_k^{m_{hij}} Y_{hijk}} - r\right\}^2.$$

Let $\psi_{hijk} = Y_{hijk}(r_{hijk} - r)$, and $Y' = \sum_h^L \frac{1}{t_h}\sum_i^{S_h}\sum_j^{1}\sum_k^{m_{hij}} Y_{hijk}$. Then, setting

$$(31) \qquad \theta = \sum_h^I \frac{1}{t_h}\sum_i^{S_h}\sum_j^{1}\sum_k^{m_{hij}} \psi_{hijk}/E\left(\sum_h^L \frac{1}{t_h}\sum_i^{S_h}\sum_j^{1}\sum_k^{m_{hij}} Y_{hijk}\right) = \frac{Y'}{EY'}\left(\frac{X'}{Y'} - r\right)$$

$$E\theta^2 = EY'^2(r' - r)^2/(EY')^2$$

is the first approximation to the mean square error.

Since EY' is evaluated in the same way as EX'(8), it is merely necessary to evaluate $EY'^2(r' - r)^2$, the numerator of $E\theta^2$. Now

$$EY'^2(r' - r)^2 = E\left[\sum_h \frac{1}{t_h}\sum_i^{S_h}\sum_j^{1}\sum_k^{m_{hij}} \psi_{hijk}\right]^2$$

$$= E\sum_h \frac{1}{t_h^2}\psi_h'^2 + E\sum_{\substack{h,q \\ h\neq q}} \frac{\psi_h'}{t_h}\frac{\psi_q'}{t_q}$$

where $\psi_h' = \sum_i^{S_h}\sum_j^{1}\sum_k^{m_{hij}} \psi_{hijk} = \sum_i^{S_h} \psi_{hi}'$.

Since $E\sum \frac{1}{t_h^2}\psi_h'^2 = E\sum_h\sum_i \psi_{hi}'^2/t_h^2 + E\sum_h \sum_{\substack{i,r\\ i\neq r}} \psi_{hi}'\psi_{hr}'/t_h^2$

$$EY'^2(r'-r)^2 = E\sum_h \frac{1}{t_h^2}\sum_i \psi_{hi}'^2 + E\sum_h \frac{1}{t_h^2}\sum_{\substack{i,r\\ i\neq r}} \psi_{hi}'\psi_{hr}' + E\sum_{\substack{h,q\\ h\neq q}} \frac{\psi_h'}{t_h}\frac{\psi_q'}{t_q}. \tag{32}$$

The first term in the right-hand side of (32) is

$$E\sum_{h,i}\psi_{hi}'^2\frac{1}{t_h^2} = \sum_{h,i,j}\frac{1}{t_h^2}\frac{P_{hj}}{P_h}\frac{m_{hij}}{M_{hij}}\frac{M_{hij}-m_{hij}}{M_{hij}-1}\sum_k \psi_{hijk}^2 + \sum_{h,i,j}\frac{1}{t_h^2}\frac{P_{hj}}{P_h}\frac{m_{hij}}{M_{hij}}\frac{m_{hij}-1}{M_{hij}-1}\Big(\sum_k \psi_{hijk}\Big)^2. \tag{33}$$

The second term of (32) is

$$E\sum_h \frac{1}{t_h^2}\sum_{\substack{i,r\\ i\neq r}}\psi_{hi}'\psi_{hr}' = \sum_{h,j}\frac{1}{t_h^2}\Big(\sum_i \frac{m_{hij}}{M_{hij}}\psi_{hij}\Big)^2 - \sum_{h,j}\frac{1}{t_h^2}\frac{P_{hj}}{P_h}\sum_i \frac{m_{hij}^2}{M_{hij}^2}\psi_{hij}^2 \tag{34}$$

where

$$\psi_{hij} = \sum_k \psi_{hijk};$$

and the third term of (32) is

$$E\sum_{\substack{h,q\\ h\neq q}}\frac{\psi_h'}{t_h}\frac{\psi_q'}{t_q} = \Big[\sum_{h,i,j}\frac{1}{t_h}\frac{P_{hj}}{P_h}\frac{m_{hij}}{M_{hij}}\psi_{hij}\Big]^2 - \sum\frac{1}{t_h^2}\Big[\sum_{i,j}\frac{P_{hj}}{P_h}\frac{m_{hij}}{M_{hij}}\psi_{hij}\Big]^2. \tag{35}$$

Therefore $EY'^2(r'-r)^2 = (33) + (34) + (35)$, and when $Y_{hijk}(r_{hijk}-r)$ is substituted for ψ_{hijk}, we have

$$\begin{aligned} EY'^2(r'-r)^2 = {} & \sum\frac{1}{t_h^2}\Big[\sum_{i,j}\frac{P_{hj}}{P_h}\frac{m_{hij}}{M_{hij}}\frac{M_{hij}-m_{hij}}{M_{hij}-1}\sum Y_{hijk}^2(r_{hijk}-r)^2 \\ & + \sum_{i,j}\frac{P_{hj}}{P_h}\frac{m_{hij}}{M_{hij}}\frac{m_{hij}-1}{M_{hij}-1}\Big[\sum_k Y_{hijk}(r_{hijk}-r)\Big]^2 \\ & + \sum_j\frac{P_{hj}}{P_h}\Big[\sum_i \frac{m_{hij}}{M_{hij}}Y_{hij}(r_{hij}-r)\Big]^2 \\ & - \sum_{i,j}\frac{P_{hj}}{P_h}\frac{m_{hij}^2}{M_{hij}^2}Y_{hij}^2(r_{hij}-r)^2 \\ & - \Big\{\sum_{i,j}\frac{P_{hj}}{P_h}\frac{m_{hij}}{M_{hij}}(r_{hij}-r)Y_{hij}\Big\}^2\Big] \\ & + \Big[\sum_{h,i,j}\frac{1}{t_h}\frac{P_{hj}}{P_h}\frac{m_{hij}}{M_{hij}}Y_{hij}(r_{hij}-r)\Big]^2. \end{aligned} \tag{36}$$

By substituting $(r_{hsjk} - r_{hsj} + r_{hsj} - r)^2$ for $(r_{hsjk} - r)^2$ in the first term of (36) and $P_{hs}M_{hsj}/P_{hsj}m_{hsj}$ for $1/l_h$ in the 1st, 2nd, and 4th terms, the sum of these three terms becomes

$$\begin{aligned}(37)\qquad &\sum_{h,s,j,k}\frac{P_{hj}}{P_h}\frac{M_{hsj}}{m_{hsj}}\frac{P_{hs}^2}{P_{hsj}^2}F_{hsj}Y_{hsjk}^2(r_{hsjk}-r_{hsj})^2\\ &+2\sum_{h,s,j,k}\frac{P_{hj}}{P_h}\frac{M_{hsj}}{m_{hsj}}\frac{P_{hs}^2}{P_{hsj}^2}F_{hsj}Y_{hsjk}^2(r_{hsjk}-r_{hsj})(r_{hsj}-r)\\ &+\sum_{h,s,j}\frac{P_{hj}}{P_h}\frac{M_{hsj}}{m_{hsj}}\frac{P_{hs}^2}{P_{hsj}^2}F_{hsj}(r_{hsj}-r)^2\left[\sum_k Y_{hsjk}^2-\frac{Y_{hsj}^2}{M_{hsj}}\right]\end{aligned}$$

where $F_{hsj} = (M_{hsj} - m_{hsj})/(M_{hsj} - 1)$ and $r_{hsj} = \sum_k X_{hsjk}/\sum_k Y_{hsjk}$.

When we substitute the appropriate value for $1/l_h$ in the 3rd, 5th, and 6th terms of (36), the sum of these terms becomes

$$\begin{aligned}(38)\qquad &\sum_{h,j}\frac{P_{hj}}{P_h}\left[\sum_s\frac{P_{hs}}{P_{hsj}}Y_{hsj}(r_{hsj}-r)\right]^2-\sum_h\left[\sum_{s,j}\frac{P_{hj}}{P_h}\frac{P_{hs}}{P_{hsj}}Y_{hsj}(r_{hsj}-r)\right]^2\\ &+\left[\sum_{h,s,j}\frac{P_{hj}}{P_h}\frac{P_{hs}}{P_{hsj}}Y_{hsj}(r_{hsj}-r)\right]^2.\end{aligned}$$

Now

$$(39)\qquad \sum_s\frac{P_{hs}}{P_{hsj}}Y_{hsj}(r_{hsj}-r)=\sum_s P_{hs}\left(\frac{X_{hsj}}{P_{hsj}}-\frac{Y_{hsj}}{P_{hsj}}r\right)=P_h(R_{hj(A)}-rR_{hj(A):Y})$$
$$=P_hR_{hj(A):Y}(\bar{r}_{hj(A)}-r)$$

where $\bar{r}_{hj(A)} = R_{hj(A)}/R_{hj(A):Y}$, and

$$(40)\qquad \sum_{s,j}\frac{P_{hj}}{P_h}\frac{P_{hs}}{P_{hsj}}Y_{hsj}(r_{hsj}-r)=\sum_j P_{hj}(R_{hj(A)}-rR_{hj(A):Y})=P_h(R_{h(A)}-rR_{h(A):Y})$$
$$=P_hR_{h(A):Y}(\bar{r}_{h(A)}-r)$$

where $\bar{r}_{h(A)} = R_{h(A)}/R_{h(A):Y}$.

Substituting (39) and (40) in (38), we have

$$\begin{aligned}(41)\qquad &\sum_{h,j}(P_{hj}/P_h)P_h^2R_{hj(A):Y}^2(\bar{r}_{hj(A)}-r)^2-\sum_h P_h^2R_{h(A)\cdot Y}^2(\bar{r}_{h(A)}-r)^2\\ &+[\sum_h P_hR_{h(A):Y}(\bar{r}_{h(A)}-r)]^2.\end{aligned}$$

By substituting $(\bar{r}_{hj(A)} - \bar{r}_{h(A)} + \bar{r}_{h(A)} - r)^2$ for $(\bar{r}_{hj(A)} - r)^2$ in the first term in (41) and expanding, (41) becomes

$$\begin{aligned}(42)\qquad &\sum_{h,j}P_h^2\frac{P_{hj}}{P_h}R_{hj(A)\cdot Y}^2(\bar{r}_{hj(A)}-\bar{r}_{h(A)})^2+2\sum_{h,j}P_h^2\frac{P_{hj}}{P_h}R_{hj(A)\cdot Y}^2(\bar{r}_{hj(A)}-\bar{r}_{h(A)})(\bar{r}_{h(A)}-r)\\ &+\sum_h P_h^2(\bar{r}_{h(A)}-r)^2\left[\sum_j\frac{P_{hj}}{P_h}R_{hj(A):Y}^2-R_{h(A)\cdot Y}^2\right]+[\sum_h P_hR_{h(A):Y}(\bar{r}_{h(A)}-r)]^2.\end{aligned}$$

Hence, since $(EY')^2\, E\theta^2 = (37) + (42)$,

$$
\begin{aligned}
(EY')^2 E\theta^2 = {} & \sum_{h,i,j} P_{hi}^2 \frac{P_{hj}}{P_h} \frac{M_{hij} - m_{hij}}{M_{hij} - 1} \frac{\sum_k Y_{hijk}^2 (r_{hijk} - r_{hij})^2}{m_{hij} M_{hij} \overline{P}_{hij}^2} \\
& + 2 \sum_{h,i,j} P_{hi}^2 \frac{P_{hj}}{P_h} \frac{M_{hij} - m_{hij}}{M_{hij} - 1} \frac{\sum_k Y_{hijk}^2 (r_{hijk} - r_{hij})(r_{hij} - r)}{m_{hij} M_{hij} \overline{P}_{hij}^2} \\
& + \sum_{h,i,j} P_{hi}^2 \frac{P_{hj}}{P_h} \frac{M_{hij} - m_{hij}}{M_{hij} - 1} \sigma_{hij:Y}^2 \frac{(r_{hij} - r)^2}{m_{hij} \overline{P}_{hij}^2} \\
& + \sum_{h,j} P_h^2 (P_{hj}/P_h) R_{hj(A)\cdot Y}^2 (\bar{r}_{hj(A)} - \bar{r}_{h(A)})^2 \\
& + 2 \sum_{h,j} P_h^2 (P_{hj}/P_h) R_{hj(A)\cdot Y}^2 (\bar{r}_{hj(A)} - \bar{r}_{h(A)})(\bar{r}_{h(A)} - r) \\
& + \sum_{h,j} P_h^2 (\bar{r}_{h(A)} - r)^2 (P_{hj}/P_h)(R_{hj(A).Y} - R_{h(A).Y})^2 \\
& + [\sum_h P_h R_{h(A):Y} (\bar{r}_{h(A)} - r)]^2
\end{aligned}
\tag{43}
$$

where $\sigma_{hij\,Y}^2 = \sum_k^{M_{hij}} (Y_{hijk} - \bar{Y}_{hij})^2 / M_{hij}$ and $\bar{Y}_{hij} = \sum_k^{M_{hij}} Y_{hijk}/M_{hij} = Y_{hij}/M_{hij}$.

The approximation to $E(r' - r)^2$ is given by (43) divided by $(EY')^2$. By ignoring the 2nd, 5th, and 7th terms which are negligible for a large class of populations, we obtain (13).

The variance of X' is derived from (43) by simply substituting $\overline{P}_{hij}/P$ for Y_{hijk} in (43). This follows from the considerations given below:

Since $r' = X'/Y'$, and X' is the numerator of r', $\sigma_{X'}^2$ is given by $\sigma_{r'}^2$ when the denominator, Y', is identically equal to unity in repeated samplings.

$$\text{Since } \frac{1}{t_h} = \frac{M_{hij} P_{hi}}{m_{hij} \overline{P}_{hij}} = \frac{P_{hi}}{m_{hij} \overline{P}_{hij}} \text{ from (5),}$$

the denominator of r' which is equal to

$\sum_h^L \sum_i^{S_h} \sum_j^1 \sum_k^{m_{hij}} \frac{P_{hi}}{m_{hij} \overline{P}_{hij}} Y_{hijk}$, will be identically equal to unity in repeated sampling when Y_{hijk} is set equal to $\overline{P}_{hij}/P$ where $P = \Sigma P_h$.

The formula for the mean square error of X' (9), of course is exact since the error term

$$E\{Y'^2/(EY')^2\}\{r' - r\}^2 = 0.$$

It may be pointed out that $\sigma_{X'}^2$ may be obtained directly and more simply without the use of (29) since X' is not estimated from the ratio of random variables.

From (29), the error term for the approximation to $E(r' - r)^2$, $(43)/(EY')^2$, is given by $E\left(1 - \frac{Y'^2}{(EY')^2}\right)\{r' - r\}^2$. This cannot be expressed as a simple func-

tion of the individual observations, but useful maxima and minima for it may be obtained. A method for obtaining the upper and lower bounds of the variance of r' is simply attained from the following inequalities which hold independent of the joint distribution of X' and Y'.

$$\frac{EY'^2}{Y_{\max}^2}(r' - r)^2 \leq E(r' - r)^2 \leq \frac{EY'^2}{Y_{\min}^2}(r' - r)^2 \tag{44}$$

where $Y_{\max}$ is the maximum value of the Y' obtained simply by choosing or estimating the largest Y_h' for each stratum. $Y_{\min}$ (the minimum value of Y') is obtained in a similar manner.

Eq. 44 when evaluated turns out to be

$$\frac{(EY')^2 E\theta^2}{Y_{\max}^2} \leq E(r' - r)^2 \leq \frac{(EY')^2 E\theta^2}{Y_{\min}^2} \tag{45}$$

where $(EY')^2 E\theta^2$ is given by (43).

Eq. (45) will serve adequately as an indicator of the accuracy of $E\theta^2$ for sampling systems in which the variability of the Y's within strata is restricted. However, in other designs, where stratification is not used and the variability in the Y's is not restricted the limits given by (45) may be too broad to be useful.

REFERENCES

[1] A. L. Bowley, "Measurement of the precision attained in sampling," *Bulletin de L'Institut International de Statistique* Tome XXII, (1926), 1 ere Livraison.

[2] W. G. Cochran, "The use of analysis of variance in enumeration by sampling," *Jour. Amer. Stat Assoc*, Vol 34 (1939), pp. 492–510.

[3] W. G. Cochran, "Sampling theory when the sampling units are of unequal sizes," *Jour. Amer Stat Assoc.*, Vol. 37 (1942), pp. 199–212.

[4] Calvert L. Dedrick and Morris H. Hansen, *Final Report on Total and Partial Unemployment, Vol IV. The Enumerative Check Census*, U. S. Govt. Printing Office (1938)

[5] Morris H. Hansen and William N. Hurwitz, "Relative efficiencies of various sampling units in population inquiries," *Jour Amer Stat Assoc.*, Vol 37 (1942). pp. 89–94.

[6] R. J. Jessen, *Statistical investigation of a sample survey for obtaining farm facts*, Research Bulletin 304 (1942), Iowa State College, Ames, Iowa

[7] P. C. Mahalanobis, "A sample survey of the acreage under jute in Bengal," *Sankhyā*, Vol. 4 (1940), pp 511–530.

[8] J. Neyman, "On the two different aspects of the representative method, a method of stratified sampling and the method of purposive selection," *Jour Royal Stat. Soc*, New Series, Vol. 97 (1934), pp 558–606.

[9] J. Neyman, "Contribution to the theory of sampling human populations," *Jour. Amer. Stat. Assoc.*, Vol 35 (1938), pp 101–116.

[10] F. Yates and I. Zacopanay, "The estimation of the efficiency of sampling, with special reference to sampling for yield in cereal experiments." *Jour. Agr. Sci*, Vol. 25 (1935), pp. 543–577

[11] Bureau of the Census, "A revised sample for current surveys," February 1943

MULTIPLE SAMPLING WITH CONSTANT PROBABILITY

By Walter Bartky

The University of Chicago

1. Introduction. In an attempt to reduce inspection costs, manufacturers have frequently resorted to sampling procedure in which the disposition of an aggregate or lot of similar units does not necessarily depend upon the results of a single sample. In practice, however, the number of permissible additional samples is limited to one or two; nevertheless, if the lot is very large, an appreciable reduction in the expected sample may be accomplished by allowing a greater number of additional samples. In this article probability formulae will be derived for an inspection procedure for infinite lots in which the number of additional samples is not limited and may be any number depending upon the results of the sampling This development will be limited to the simple case of attribute inspection in which the units fall into two categories—satisfactory units or defective units. If p denotes the fraction defective in an infinite lot, then the probability of finding exactly m defective units or defects in a sample of n is

$$P(m, n) = \binom{n}{m} p^m q^{n-m}, \qquad q = 1 - p. \tag{1}$$

Since $P(m, n)$ is the probability of m successes in n trials with constant probability of success p, though the terminology of commercial inspection will be used in this article, the results are applicable to other situations involving repeated trials with constant probability of success.

In contrast with multiple sampling, a single sample inspection procedure for lots of the type here considered is one in which a lot of units is accepted or rejected on the basis of the number of defective units found in the sample. Thus a lot is accepted if the number of defects is at most an integer c the "acceptance number," and rejected if the number exceeds c. For an infinite lot containing a fraction p of defects and a sample of n units, the probability of accepting is by (1)

$$\prod_s (c, n) = \sum_{m \leq c} P(m, n), \tag{2}$$

and the probability for rejection is the difference between this sum and unity.

2. Multiple sampling. The procedure in multiple sampling is to examine first an initial sample of n_0 units If the number of defects in this initial sample is at most c the lot is accepted and if the number of defects exceeds $c + k$ (k an integer) the lot is rejected. But if the number of defects is greater than c and less than $c + k + 1$ an additional sample is removed and examined In the latter case similar criteria determine whether the lot is to be accepted or rejected or this method of sampling continued. With an infinite lot this scheme of samp-

ling has an infinite variety of forms but there are certain advantages in limiting this discussion to the following type of multiple sampling procedure.

I. *Sample Sizes:* The initial sample is of n_0 units but all additional samples are of the same size, namely n units.

II. *Condition for Acceptance:* The lot is accepted if the number of defects in initial sample of n_0 units is at most c or if after taking r additional samples of n the total number of defects in the $n_0 + rn$ units examined equals $c + r$.

III. *Condition for Rejection:* The lot is rejected if the number of defects in initial sample of n_0 units exceeds $c + k$ or if after taking r additional samples of n the total number of defects exceeds $c + r + k$.

IV. *Condition for an Additional Sample:* An additional sample of n is taken only if neither condition II nor condition III is realized.

Thus in this sampling scheme the level for acceptance as well as the level for rejection increases by unity for each additional sample of n. If at the r-*th* additional sample a lot is neither accepted nor rejected then the total number of defects in initial plus additional samples must equal one of the k numbers

$$c + r + 1,\ c + r + 2,\ \cdots,\ c + r + k.$$

Denote the probabilities for obtaining these numbers by

$$P_1(r),\ P_2(r),\ \cdots,\ P_k(r) \tag{3}$$

respectively, the subscript indicating the number of defects in excess of the acceptance level

To be accepted on the $(r + 1)$-*st* additional sample, (a) no defect must be found in the $(r + 1)$-*st* additional sample and (b) a total of $c + r + 1$ defects must be found in previous samples. The probability of (a) is given by (1), taking m equal to zero, and the probability of (b) is the first one in the set (3). Consequently the probability of accepting a lot on the $(r + 1)$-*st* additional sample is

$$P_0(r + 1) = q^n P_1(r).$$

If Π denotes the probability of eventually accepting the lot

$$\Pi = \sum_{m \leqq c} P(m, n_0) + q^n[P_1(0) + P_1(1) + P_1(2) + \cdots], \tag{4}$$

where the first term on the right is the probability of accepting on the initial sample and may be evaluated by means of (1). Furthermore

$$P_i(0) = P(c + i, n_0) \tag{5}$$

and is by (1) the probability of finding $c + i$ defects in initial sample

According to the notation (3) the probability of finding a total of $c + r + 1 + i$ defects in initial plus $r + 1$ additional samples, that is i more defects than the acceptance level, is $P_i(r + 1)$. These probabilities may be expressed as

linear combinations of the probabilities (3) with coefficients that are probabilities of the type (1). Thus

$$P_i(r+1) = \sum_j P(i-j+1, n)P_j(r) \tag{6}$$

where the sum may be made to extend for $j = 1, 2, \cdots, k$, provided one defines (1) as equal to zero for negative m. By repeated application of this linear transformation it is possible to express the probabilities (3) for additional samples in terms of the probabilities (5) for the initial sample. Thus if M denotes the $k \times k$ square matrix with elements

$$M_{ij} = P(i-j+1, n) \qquad (i, j = 1, \cdots, k), \tag{7}$$

by omitting subscripts and regarding $P(r)$ as a vector with elements given by (3), the linear transformation may be written

$$P(r+1) = MP(r). \tag{8}$$

Hence by repeated application of (8)

$$P(r) = M^r P(0) \qquad (r = 0, 1, 2, \cdots) \tag{9}$$

provided the zero power of the matrix M is defined as the identity matrix I.

The probability $P_i(r)$ cannot exceed the probability of finding exactly $c + r + i$ defects in a single sample of $n_0 + rn$ units, that is, in the notation of (1), the probability $P(c + r + i, n_0 + rn)$. Since the latter probabilities approach zero as r approaches infinity it follows that the limit of the elements of $P(r)$ as r approaches infinity is zero. Thus with this multiple sampling procedure a lot is eventually either accepted or rejected. Furthermore since the matrix M contains no negative elements and $P(0)$ may be chosen with all positive elements it follows that the elements of M^r approach zero as r approaches infinity or

$$\lim_{r\to\infty} M^r = \text{"0"}, \quad \text{the zero matrix.} \tag{10}$$

It can be demonstrated that since the limit (10) is the zero matrix the sum of the infinite geometrical series in the matrix M

$$I + M + M^2 + \cdots = (I - M)^{-1}, \tag{11}$$

where the right member is the reciprocal of the matrix $I - M$. Consequently the infinite sum of vectors

$$V = \sum_{r=0}^{\infty} P(r) = (I - M)^{-1} P(0). \tag{12}$$

This infinite sum of vectors has elements $V_1, V_2, \cdots, V_k$ of which the first element is the sum in brackets occurring in the right member of (4). Hence the probability of eventually accepting the lot

$$\prod = \sum_{m \leqq c} P(m, n_0) + q^n V_1, \tag{13}$$

and is thus by (12) and (5) expressible in terms of probabilities for the initial sample, equations (1), and the reciprocal of the matrix $I - M$.

In addition to the probability for acceptance one is also interested in the expected number, E, of additional samples. Since

$$\sum_i P_i(r-1) \qquad (r = 1, 2, 3, \cdots),$$

where the sum extends over all $i = 1, 2, \cdots, k$ is the probability of continuing to the r-th sample, it follows that

$$\sum_i P_i(r-1) - \sum_i P_i(r)$$

is the probability that lot will be either accepted or rejected on the r-th sample. Therefore the expected number of additional samples

$$\begin{aligned} E &= \sum_{r>0} r\Big[\sum_i P_i(r-1) - \sum_i P_i(r)\Big] \\ &= \sum_{r\geq 0} \sum_i P_i(r), \end{aligned}$$

or, on interchanging the order of summation and applying (12),

$$E = \sum_i V_i . \tag{14}$$

That is, the expected number of additional samples equals the sum of the elements of the vector V.

Though it is possible to develop a general expression for the reciprocal matrix $I - M$, to determine the acceptance probability, Π, as well as the expected number of additional samples it is only necessary to evaluate V. Now by (12) this vector is the solution of the linear system of equations

$$(I - M)V = P(0). \tag{15}$$

Though for k small this system could be solved directly, in order to find a form of the solution applicable for any value of k, let the expansion in power series in x of

$$[(px + q)^n - x]^{-1} = g_1 + g_2 x + g_3 x^3 + \cdots , \tag{16}$$

where the coefficients, g, are functions of p and q. On clearing of fractions and equating coefficients of like powers of x it is found that

$$g_1 = q^{-n} \tag{17}$$

and, by equating the coefficients of the first k powers of x and using the notation (7),

$$g_i - \sum_{j=1 \cdots k} M_{ij} g_j = \begin{cases} 0 & (i = 1, 2, \cdots, k-1), \\ g_{k+1} & (i = k). \end{cases} \tag{18}$$

Similarly, if the expansion in power series of

$$\frac{\sum_i P_i(0)x^i}{(px+q)^n - x} = h_1 + h_2 x + h_3 x^3 + \cdots, \tag{19}$$

where the sum is for all $i = 1, \cdots, k$, then by clearing of fractions and equating coefficients of like powers of x it is found that

$$h_1 = 0, \tag{20}$$

and

$$h_i - \sum_{j=1\cdots k} M_{ij} h_j = \begin{cases} -P_i(0) & (i = 1, \cdots, k-1), \\ -P_k(0) + h_{k+1} & (i = k). \end{cases} \tag{21}$$

It follows from equations (18) and (21) that if

$$V_i = g_i h_{k+1}/g_{k+1} - h_i \qquad (i = 1, \cdots, k), \tag{22}$$

then V, the vector with these elements, will satisfy equation (15). Since by (17) and (20)

$$V_1 = q^{-n} h_{k+1}/g_{k+1}, \tag{23}$$

the probability for eventually accepting the lot is by (13) expressible as

$$\Pi = \sum_{m \leqq c} P(m, n_0) + h_{k+1}/g_{k+1}, \tag{24}$$

while the expected number of additional samples is the sum of elements (22) of V_i.

These results will now be summarized and simplified formulae derived for special cases. In the summary all probabilities are expressed by means of (5) in terms of the probabilities (1).

3. Summary of multiple sampling formulas. For this multiple sampling procedure the initial sample is n_0 and the additional samples are n. A lot is accepted if on the r-th additional sample the total number of defects found is at most $c + r$ and rejected if the total exceeds $c + r$. An infinite lot containing a fraction p of defects is either accepted or rejected, the probability of acceptance being given by

$$\Pi = \sum_{m \leqq c} \binom{n_0}{m} p^m q^{n-m} + h_{k+1}/g_{k+1} \qquad (q = 1 - p), \tag{25}$$

and the probability of rejection is $1 - \Pi$. The expected number of additional samples is

$$E = \frac{h_{k+1}}{g_{k+1}} \sum_i g_i - \sum_i h_i, \tag{26}$$

where the sum extends over $i = 1, 2, \cdots, k$. The g_i and h_i are the coefficients in power series of x in the expansions of:

$$\frac{1}{(px+q)^n - x} = g_1 + g_2 x + g_3 x^2 + \cdots, \tag{27}$$

$$\sum_i \frac{\binom{n_0}{c+i} p^{c+i} q^{n_0-c-i} x^i}{(px+q)^n - x} = h_1 + h_2 x + h_3 x^2 + \cdots, \tag{28}$$

where the sum is for all $i = 1, 2, \cdots, k$. These formulae apply to all finite values of c and k provided the binomial coefficient is zero for values of the argument falling outside those occurring in the ordinary expansion of an integral power of a binomial.

4. Computation of coefficients g and h. If the denominator in (27) is first expanded in power series in

$$x(px+q)^{-n}$$

and then the resulting negative powers of binomials expanded in power series in x, it is found that

$$\begin{aligned} g_1 &= q^{-n}, \\ g_2 &= q^{-2n} - \binom{n}{1} p q^{-n-1}, \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ g_k &= q^{-kn} - \sum_{m=1,\cdots,k-1} (-1)^{m+1} \binom{(k-m)n+m-1}{m} \\ &\qquad \times p^m q^{-kn+mn-m}, \qquad k \neq 1. \end{aligned} \tag{29}$$

By (28) the coefficients h are expressible in terms of the g's,

$$\begin{aligned} h_1 &= 0, \\ h_k &= \sum_{i=1,\cdots,k-1} \binom{n_0}{c+i} p^{c+i} q^{n_0-c-i} g_{k-i}, \qquad k \neq 1. \end{aligned} \tag{30}$$

Other expressions for the coefficients may be derived from the theory of functions of a complex variable. Thus by Cauchy's Integral Formula

$$\begin{aligned} g_{k+1} &= \frac{1}{2\pi\sqrt{-1}} \int_C \frac{dx}{x^{k+1}[(px+q)^n - x]}, \\ h_{k+1} &= \frac{1}{2\pi\sqrt{-1}} \int_C \frac{S(x)\,dx}{x^{k+1}[(px+q)^n - x]}, \end{aligned} \tag{31}$$

where

$$S(x) = \sum_{i=1,\cdots,k} \binom{n_0}{c+i} p^{c+i} q^{n_0-c-i} x^i, \tag{32}$$

and the closed path of integration C in the complex plane only includes the pole at the origin. Since the integrands are rational functions and the point at infinity is not a singularity for either integrand, these integrals taken about the origin are equal to the negative sum of the corresponding integrals taken about the zeros of

$$(px + q)^n - x.$$

If $p \neq n^{-1}$ it can be demonstrated that there are n distinct zeros $x_1, x_2, \cdots, x_n$ corresponding to the solutions of the algebraic equation

$$(px_s + q)^n = x_s \qquad (s = 1, \cdots, n). \tag{33}$$

One solution is obviously

$$x_1 = 1, \tag{34}$$

and for $p = n^{-1}$ this solution is a double root

The integrals about these zeros are obtainable from Cauchy's Integral Formula and after integrating and simplifying the resulting sum by means of (33) it is found that for the case $p \neq n^{-1}$,

$$\begin{aligned} g_{k+1} &= \frac{1}{1-np} + \sum_{s=2,\cdots,n} \frac{px_s + q}{x_s^{k+1}[q - (n-1)px_s]}, \\ h_{k+1} &= \frac{S(1)}{1-np} + \sum_{s=2,\cdots,n} \frac{(px_s + q)S(x_s)}{x_s^{k+1}[q - (n-1)px_s]}. \end{aligned} \tag{35}$$

If the power series (27) is multiplied by the series

$$(1-x)^{-1} = 1 + x + x^2 + x^3 + \cdots,$$

the resulting product

$$\frac{1}{(1-x)[(px+q)^n - x]} = g_1 + (g_1 + g_2)x + (g_1 + g_2 + g_3)x^2 + \cdots,$$

so that, by Cauchy's Integral Formula,

$$G_k = \sum_{i=1,\cdots,k} g_i = \frac{1}{2\pi\sqrt{-1}} \int_C \frac{dx}{x^k(1-x)[(px+q)^n - x]}. \tag{36}$$

Similarly the sum of the coefficients h that occur in the right member of (26) may be written

$$H_k = \sum_{i=1,\cdots,k} h_i = \frac{1}{2\pi\sqrt{-1}} \int_C \frac{S(x)\,dx}{x^k(1-x)[(px+q)^n - x]}. \tag{37}$$

The integrals (36) and (37) are of the same type as (31), and by employing the same method of integrating used in deriving (35), the following expressions for the sums of coefficients g and h occurring in (26) are obtained:

$$G_k = \sum_i g_i = \frac{k}{1-np} - \frac{n(n-1)p^2}{2(1-np)^2} + \sum_{s=2,\cdots,n} \frac{px_s + q}{x_s^k(1-x_s)[q-(n-1)px_s]},$$

$$(38)\quad H_k = \sum_i h_i = \frac{kS(1) - S'(1)}{1-np} - \frac{n(n-1)p^2 S(1)}{2(1-np)^2} + \sum_{s=2,\cdots,n} \frac{(px_s+q)S(x_s)}{x_s^k(1-x_s)[q-(n-1)px_s]}$$

provided $np \neq 1$. Here $S'(1)$ is the derivative of (32) with respect to x evaluated for $x = 1$. For the special case $np = 1$, two of the roots of (33)

$$x_1 = x_2 = 1,$$

and the integrals (36), (37) and (38) become respectively

$$(n-1)g_{k+1} = 2kn + \tfrac{8}{3}n - \tfrac{4}{3} + \sum_{s\geqq 3} \frac{n + x_s - 1}{x_s^{k+1}(1-x_s)},$$

$$(n-1)h_{k+1} = (2kn + \tfrac{8}{3}n - \tfrac{4}{3})S(1) - 2nS'(1) + \sum_{s\geqq 3} \frac{n + x_s - 1}{x_s^{k+1}(1-x_s)} S(x_s),$$

$$(39)\quad (n-1)\sum_i g_i = k^2 n + \tfrac{5}{3}kn + \tfrac{1}{18}n - \tfrac{4}{3}k - \tfrac{1}{18} - \tfrac{1}{9}n^{-1} + \sum_{s\geqq 3} \frac{n + x_s - 1}{x_s^k(1-x_s)^2},$$

$$(n-)1\sum_i h_i = (k^2 n + \tfrac{5}{3}kn + \tfrac{1}{18}n - \tfrac{4}{3}k - \tfrac{1}{18} - \tfrac{1}{9}n^{-1})S(1) - (\tfrac{4}{3}n - \tfrac{4}{3} + 2kn)S'(1) + nS''(1) + \sum_{s\geqq 3} \frac{n + x_s - 1}{x_s^k(1-x_s)^2} S(x_s),$$

where the sum extends over all roots of (33) that are not equal to unity. Here $S'(1)$ and $S''(1)$ are the first and second derivatives of (32) with respect to x for $x = 1$ and $p = n^{-1}$.

Formulas (35), (38) and (39) require for their evaluation the solutions of equation (33). For n greater than unity there are just two positive real solutions, say $x_1 = 1$ and x_2. If n is even all other roots are complex numbers, while if n is odd they are complex with the exception of one negative real root. Consequently by (33) for $s = 3, 4, \cdots, n$ the absolute values of the roots satisfy the inequality

$$(p \mid x_s \mid + q)^n > x_s,$$

and consequently the $\mid x_s \mid$ cannot be between x_1 and x_2. But equation (33) may be written

$$\frac{(px_s + q)^n - 1}{(px_s + q) - 1} = \frac{1}{p} \qquad (s \neq 1)$$

so that for $s = 2, 3, \cdots, n$

$$\sum_{i} (px_s + q)^i = 1/p \tag{40}$$

where the sum is taken for $i = 0, 1, \cdots, n-1$ and therefore

$$\sum_{i} (p\,|\,x_s\,| + q)^i > 1/p \qquad (s = 3, 4, \cdots, n).$$

Now x_2 is the only real and positive solution of (40), consequently, in order to satisfy the inequality, the absolute values of roots corresponding to $s = 3, 4, \cdots, n$ must exceed x_2. On combining this result with the former, it follows that

$$|\,x_s\,| > 1 \quad \text{and} \quad x_2. \tag{41}$$

Consequently for large values of k the most important terms in the right members of (35), (38) and (39) correspond to the real positive roots $x_1 = 1$ and x_2 of equation (33). By omitting the terms corresponding to $s = 3, \cdots, n$ one can derive approximations to the g and h and their sums applicable for large k values. In fact for np near unity the roots corresponding to $s = 3, 4, \cdots$ are considerably greater than unity as is illustrated in the following table of roots for the case $np = 1$:

$n = 2, \quad p = 1/2; \quad x_s = 1, 1;$

$n = 3, \quad p = 1/3; \quad x_s = 1, 1, -8;$

$n = 4, \quad p = 1/4; \quad x_s = 1, 1, -7 \pm 4\sqrt{-2};$

$n = 5, \quad p = 1/5; \quad x_s = 1, 1, -12.2531 \cdots,$

$$-4.8734 \cdots \pm 7.7343 \cdots \sqrt{-1}$$

and for $s = 3, 4, 5, \cdots, |\,x_s\,|$ is greater than 8.

For very large values of n and small values of p one can find approximate values for the roots by solving the limit equation obtained from (33) by putting

$$a = np$$

and letting n approach infinity. This equation is

$$e^{a(x_s - 1)} = x_s, \tag{42}$$

where e is the base of the natural logarithms. For the case $a = 1$, the roots are $1, 1, 3.0891 \cdots \pm 7.4602 \cdots \sqrt{-1}$, $3.66 \cdots \pm 13.88 \cdots \sqrt{-1}$ and

$$x_s = \frac{b(1 + \log_e b)}{b^2 + 1}(b - \sqrt{-1}) + b\sqrt{-1} \quad \text{approximately,}$$

where

$$b = (2u + 1/2)\pi, \qquad u = 4, 5, 6, \cdots.$$

From equation (39) and these numerical results it follows that even with k as small as 3 the percentage error for the case $np = 1$ introduced in g_4 by omitting the terms in the indicated sum is less than .002%. Consequently for all practical purposes one may omit the complex and negative roots for values of k greater than 3 in computing the g's for np in the neighborhood of unity For smaller values of k the exact values of the g's are readily obtainable from (29).

5. Special cases. Consider first the case in which $c < 0$ and $n_0 \leqq k + c$. With these conditions, under no circumstances could a lot be accepted or rejected on the initial sample and the indicated sum in the right member of (25) is zero. Furthermore for this case the sum (32) becomes

$$S(x) = (px + q)^{n_0}x^{-c}. \tag{43}$$

Consequently it follows from (33) that

$$S(x_s) = x_s^{t-c}, \tag{44}$$

where

$$t = n_0/n. \tag{45}$$

It should be noted however, that for t not an integer the right member of (44) is multiple valued and one must take that value for which

$$x_s^t = (px_s + q)^{n_0}. \tag{46}$$

Thus for real positive values of x_s, the right member of (44) is real. For integral values of t there is of course no ambiguity in the notation.

If (44) is substituted in the second equation of (35), the resulting expression for the h coefficient is of the same form as that for the g coefficient, in fact

$$h_{k+1} = g_{k-t+c+1},$$

so that by (25) the probability for acceptance is for this case

$$\Pi = g_{k-t+c+1}/g_{k+1}. \tag{47}$$

In similar manner it follows from (43) and (46) that the sum of the h coefficients, equation (38),

$$H_k = G_{k-t+c} + t$$

and hence by (26) the expected number of additional samples

$$E = \Pi G_k - G_{k-t+c} - t. \tag{48}$$

Since the initial sample is nt units and the additional samples are all equal to n units, the expected total number of units, sampled, that is, initial plus additional samples is

$$I = n_0 + nE = n(\Pi G_k - G_{k-t+c}). \tag{49}$$

Since for this case it is impossible to accept or reject on the initial sample one could combine the initial sample with the first additional sample. In fact one can continue combining initial and additional samples and thus increasing c and t provided the new initial sample n_0 and the new c value thus obtained are such that

$$c \leqq 0, \qquad n_0 = nt \leqq k + n - 1 + c \tag{50}$$

In this process of combining samples t and c increase at the same rate and consequently formula (47), and the right member of (49) are unchanged. In other words formulas (47) and (49) may also be used under conditions (50)

It was demonstrated in Section 3 that for k sufficiently large one can omit those terms in (35) and (38) corresponding to complex or negative roots of (33). If this is done the following useful approximations for the g and G are obtained:

$$\begin{aligned} g_k &= (1 - np)^{-1} + [q - (n - 1)px]^{-1}x^{-k+(1/n)}, \\ G_k &= k(1 - np)^{-1} - \tfrac{1}{2}n(n - 1)p^2(1 - np)^{-2} \\ &\quad + [q - (n - 1)px]^{-1}(1 - x)^{-1}x^{-k+(1/n)}, \end{aligned} \tag{51}$$

provided $np \neq 1$, $k \neq 1$ and x is the real positive root of

$$(px + q)^n = x \qquad (np \neq 1) \tag{52}$$

that is not equal to unity. For $np = 1$ these approximations become by (39)

$$\begin{aligned} (n - 1)g_k &= 2kn + 2n/3 - 4/3 \\ (n - 1)G_k &= k^2n + 5kn/3 + n/18 - 4k/3 - 1/18 - n^{-1}/9, \quad k \neq 1. \end{aligned} \tag{53}$$

These formulae in conjunction with formulae (47) and (49) give quite satisfactory approximations for the probability for acceptance Π and the expected total number of units sampled even when values of the subscripts employed are as small as 3. Of course the larger the value of k in (51), (52) or (53) the better these approximations.

Now the root x of (52) is greater or less than unity depending on whether the product $a = np$ is less than or greater than unity. Consequently it follows from (47) and (51) that for $c = 0$ and t finite

$$\begin{aligned} \Pi' &= \lim_{k\to\infty} \Pi = \lim_{k\to\infty} g_{k-t+1}/g_{k+1} \\ &= 1, \quad np < 1; \\ &= x^t, \quad np > 1; \end{aligned} \tag{54}$$

while by (49) and (51) the expected total number of units sampled has the limiting value

$$I' = \lim_{k\to\infty} I = \begin{cases} nt(1 - np)^{-1}, & np < 1; \\ \infty, & np > 1. \end{cases} \tag{55}$$

But k infinite implies that under no circumstance can a lot be rejected. Consequently Π' and I' are the exact values of the probability for acceptance and the expected total sample respectively for the following sampling procedure:

The initial sample is $n_0 = nt$ and all additional samples are n The lot is accepted if on the initial sample no defects are found or if after taking r additional samples a total of exactly r defects is found.

In inspection problems p is usually small and n large so that the approximation (40) may be used to determine the real positive root x, thus

$$e^{a(x-1)} = x \qquad (a = np). \tag{56}$$

It then follows from (54) and (55) that for $np > 1$

$$\begin{aligned} \frac{-\log \Pi'}{1-x} &= n_0 p, \\ \frac{-\log x}{1-x} &= np. \end{aligned} \tag{57}$$

These relations are of course equivalent to (54) and (56). Suppose that the probability Π' and the fraction p are assigned. Then the initial sample n_0, and additional sample n, will depend on only the parameter x Consider next the problem of sampling a number of lots that fall into two categories, namely those containing a fraction p of defects and those containing a fraction p^* of defects where $p^* < p$. If in addition the sampling procedure is to be such that lots with fraction p^* of defects are eventually accepted, but lots with fraction p of defects have a small assigned probability of acceptance Π', then whatever the value of x as long as the resulting $np \geqq 1$ these conditions are satisfied. Furthermore if one insists that the expected total sample for lots containing a fraction p^*, namely by (55)

$$I'(p^*) = n_0(1 - np^*)^{-1},$$

be a minimum, then it is found that

$$x = p^*/p. \tag{58}$$

This remarkably simple result is capable of still greater generalization. By an altogether different approach to the problem the author has succeeded in proving that of all possible multiple sampling procedures, the multiple sampling method here described and defined by equations (57) and (58) gives the minimum expected inspection for the problem under consideration provided n is sufficiently large.[1]

By letting both $-c$ and k approach infinity it is possible to derive probability formulae for sampling procedure in which a lot is either rejected or the sampling continues without end. These formulae are included in Table I along with other special cases derived from previously listed general formulae.

[1] *Note:* The author has postponed publication of this proof in the hope that it might be generalized to include sampling problems involving both acceptance and rejection of a lot.

TABLE I

Notation:

n = number of units in each additional sample
n_0 = number of units in initial sample
p = fraction defective in lot
$a = np$
$q = 1 - p$
c = maximum number of defects in initial sample for acceptance
$t = n_0/n$ = ratio initial sample to additional samples
$f = c + k + 1$ = minimum number of defects in initial sample for rejection
$c + r$ = number of defects in initial plus first r additional samples for acceptance
$f + r = c + k + 1 + r$ = minimum number of defects in initial plus first r additional samples for rejection
Π = probability of eventually accepting lot with fraction p defects
$1 - \Pi$ = probability of eventually rejecting lot with fraction p defects
I = expected total number of units sampled (i.e., initial plus whatever additional samples are sampled).
x = real positive root different from unity of the equation $(px + q)^n = x$.

Conditions	Π	I
(a) $k = 1$, $c = 0$, $f = 2$	$q^{n_0} \times \dfrac{1 - (n - n_0)\,pq^{n-1}}{1 - npq^{n-1}}$	$n_0 \times \dfrac{1 - (q^{n-1} - q^{n_0-1})np}{1 - npq^{n-1}}$
(b) $k = 1$, $c = 0$, $f = 2$, $n_0 = n$	$q^n(1 - npq^{n-1})^{-1}$	$n(1 - npq^{n-1})^{-1}$
(c) $c = -k$, $f = 1$	q^{n_0-n}/g_{k+1}	$n_0 + nq^{n_0-n}G_k/g_{k+1}$
(d) $k = 1$, $c = -1$, $f = 1$	$q^{n_0+n}(1 - npq^{n-1})^{-1}$	$n_0 + nq^{n_0}(1 - npq^{n-1})^{-1}$
(e) $k = 2$, $c = -2$, $f = 1$	$\dfrac{q^{n_0+2n}}{1 - 2npq^{n-1} + \dfrac{n(n+1)}{2}p^2q^{2n-2}}$	$n_0 + \dfrac{nq^{n_0}(1 + q^n - npq^{n-1})}{1 - 2npq^{n-1} + \dfrac{n(n+1)}{2}p^2q^{2n-2}}$

TABLE I—*Concluded*

Conditions	Π	I
(f) $k = -c$ $= \infty$ $f = 1$	0 for $np > 1$ $q^{n_0-n}(1-np)$ for $np < 1$*	$n_0 + nx(1-x)^{-1}$ for $np > 1$ ∞ for $np < 1$
(g) $c = 0$ $n = 2$ $n_0 = f$ $= k+1$	$\dfrac{1}{1+(p/q)^{n_0}}$	$\dfrac{n_0(2\Pi - 1)}{q-p}$
(h) $c = 0$ $n = 2$ $n_0 = f$ $= k+1$ $p = 1/2$	0.5	n_0^2
(i) $c = 0$ $n_0 = n$	g_k/g_{k+1}	$n(\Pi G_k - G_{k-1})$
(j) $c = 0$ $k = \infty$	1 $(np < 1)$ $x^{n_0/n}$ $(np > 1)$	$n_0(1-np)^{-1}$ $(np < 1)$ ∞ $(np > 1)$

* In this sampling procedure a lot cannot be accepted so that Π is the probability that additional samples will be taken without end. The probability of rejecting lot is however $1 - \Pi$.

TABLE II

Values of g and G for Limit $n = \infty$, $p = 0$

$np =$ $a =$	0.2558	0.4024	0.6931	1.0000	1.3863	2.0118	2.5584
$x =$	10	5	2	1	.5	.2	.1
g_1	1.292	1.495	2.000	2.718	4.000	7.477	12.915
g_2	1.338	1.634	2.614	4.671	10.455	40.86	133.76
g_3	1.3432	1.665	2.935	6.667	23.48	208.2	1343.2
g_4	1.3437	1.6717	3.097	8.667	49.55	1045.	13.4×10^3
g_5	1.3438	1.6729	3.178	10.667	101.70	5228.	134×10^3
g_∞	1.3438	1.6732	3.2589	∞	∞	∞	∞
G_1	1.292	1.495	2.000	2.718	4.000	7.477	12.915
G_2	2.629	3.130	4.614	7.389	14.45	48.34	146.7
G_3	3.972	4.795	7.549	14.05	37.93	256.5	1490
G_4	5.316	6.467	10.65	22.72	87.5	1301.	14.9×10^3
G_5	6.660	8.140	13.82	33.39	189.2	6529.	149×10^3

As an illustration of the method of application of these formulae, suppose that the sampling procedure is to be such that the probability, Π, of accepting a "p" value of $0.5 + \epsilon$ equals the probability of rejecting a "p" value of $0.5 - \epsilon$. This condition on probabilities is by Table I, formula (g), always satisfied if $c = 0$, $n = 2$, and $n_0 = k + 1$. This corresponds to a multiple sampling scheme in which additional samples are only two units each and a lot is accepted or rejected on initial sample if none or all units are defective With $\epsilon = 0.1$ and $\Pi \leqq 1/6$, one can take $n_0 = 4$ and $k = 3$. The expected total number of units examined depends on "p" and varies for this numerical case from 4, for $p = 0$ or 1, to a maximum of 16, for $p = 0.5$. Nevertheless a single sample plan satisfying the same conditions would require a sample of 23 units whatever the value of p.

The previous problem is, however, not typical of those encountered in commercial inspection for in such situations p is usually very small. In practice one can generally replace the formulae in Table I by their limiting values for $n = \infty$, $p = 0$, and $np = a$. Table II gives the limiting values of the g and G as well as x for a small number of values of a

Finally the justification for multiple sampling lies in the fact that a reduction in the expected total sample is possible. Though this paper is limited to the consideration of a very elementary type of sampling, it indicates that it might be worth while to investigate the possibility of utilizing the methods of multiple sampling in inspection for variables. Unfortunately serious mathematical difficulties are even encountered in so simple a problem as multiple sampling from a normal population for the mean.

AN EXACT TEST FOR RANDOMNESS IN THE NON-PARAMETRIC CASE BASED ON SERIAL CORRELATION[1]

By A. Wald and J. Wolfowitz

Columbia University

1. Introduction. A sequence of variates $x_1, \cdots, x_N$ is said to be a random series, or to satisfy the condition of randomness, if $x_1, \cdots, x_N$ are independently distributed with the same distribution; i e., if the joint cumulative distribution function (c.d f.) of $x_1, \cdots, x_N$ is given by the product $F(x_1) \cdots F(x_N)$ where $F(x)$ may be any c.d.f.

The problem of testing randomness arises frequently in quality control of manufactured products. Suppose that x in some quality character of a product and that $x_1, x_2, \cdots, x_N$ are the values of x for N consecutive units of the product arranged in some order (usually in the order they were produced). The production process is said to be in a state of statistical control if the sequence $(x_1, \cdots, x_N)$ satisfies the condition of randomness. A number of tests of randomness have been devised for purposes of quality control, all having the following features in common: 1) They are based on runs in the sequence $x_1, \cdots, x_N$. 2) The test procedure is invariant under topologic transformation of the x-axis, i e., the test procedure leads to the same result if the original variates $x_1, \cdots, x_N$ are replaced by $x_1', \cdots, x_N'$ where $x_\alpha' = f(x_\alpha)$ and $f(t)$ is any continuous and strictly monotonic function of t. 3) The size of the critical region, i.e., the probability of rejecting the hypothesis of randomness when it is true, does not depend on the common c.d.f. $F(x)$ of the variates $x_1, \cdots, x_N$. Condition (3) is *a fortiori* fulfilled if condition (2) is satisfied and if $F(x)$ is continuous. The fulfillment of condition (3) is very desirable, since in many practical applications the form of the c.d.f. $F(x)$ is unknown

Tests of randomness are of importance also in the analysis of time series (particularly of economic time series) where they are frequently based on the so-called serial correlation. The serial correlation coefficient with lag h is defined by the expression[2] (see, for instance, Anderson [1])

$$R_h = \frac{\sum_{\alpha=1}^{N} x_\alpha x_{h+\alpha} - \left(\sum_{\alpha=1}^{N} x_\alpha\right)^2 \Big/ N}{\sum_{\alpha=1}^{N} x_\alpha^2 - \left(\sum_{\alpha=1}^{N} x_\alpha\right)^2 \Big/ N} \tag{1}$$

where $x_{h+\alpha}$ is to be replaced by $x_{h+\alpha-N}$ for all values of α for which $h + \alpha > N$. The distribution of R_h has recently been studied by R. L. Anderson [1], T. Koopmans [2], L. C. Young [3], J. v. Neumann [4, 5], B. I Hart and J. v. Neu-

[1] Presented to the Institute of Mathematical Statistics and the American Mathematical Society at a joint meeting at New Brunswick, New Jersey, on September 13, 1943

[2] Some authors (see, for instance, [2] p. 27, equation (61)) use a non-circular definition.

mann [6], and J. D. Williams [7], under the assumption that $x_1, \cdots, x_N$ are independently distributed with the same normal distribution. Thus, in addition to the randomness of the series $(x_1, \cdots, x_N)$ it is assumed that the common c.d.f. of the variates $x_1, \cdots, x_N$ is normal This is a restrictive assumption since frequently the form of the common c.d.f $F(x)$ of the variates $x_1, \cdots, x_N$ is unknown.

The purpose of this paper is to develop a test procedure based on R_h such that (*a*) if $F(x)$ is continuous the size of the critical region does not depend on the common c.d.f. $F(x)$ of the variates $x_1, \cdots, x_N$, thus making an exact test of significance possible also when nothing is known about $F(x)$ except its continuity; (*b*) if $F(x)$ is not continuous, but all its moments are finite and its variance is positive, the size of the critical region approaches, as $N \to \infty$, the value it would have if $F(x)$ were continuous. Thus in the limit an exact test is possible in this case as well. We will refer to the case where the form of $F(x)$ is unknown as the non-parametric case, in contrast to the case when it is known that $F(x)$ is a member of a finite parameter family of c.d.f.'s

The test based on the serial correlation seems to be suitable if the alternative to randomness is the existence of a trend[3] or of some regular cyclical movement in the data. In the analysis of time series it is frequently assumed that this is the case and this is perhaps the reason why tests based on serial correlation are widely used in the analysis of time series In quality control of manufactured products the existence of a trend is often considered as the alternative to randomness, caused perhaps by the steady deterioration of a machine in the production process Thus, tests of randomness based on serial correlation could also be used in quality control.

2. An exact test procedure based on R_h. Let a_α be the observed value of $x_\alpha (\alpha = 1, \cdots, N)$. Consider the subpopulation where the set $(x_1, \cdots, x_N)$ is restricted to permutations of $a_1, \cdots, a_N$. In this subpopulation the probability that $(x_1, \cdots, x_N)$ is any particular permutation $(a_1', \cdots, a_N')$ of $(a_1, \cdots, a_N)$ is equal to $1/N!$ if the hypothesis to be tested, i.e., that of randomness, is true. (If two of the a_i $(i = 1, 2, \cdots, N)$ are identical we assume that some distinguishing index is attached to each so that they can then be regarded as distinct and so that there still are $N!$ permutations of the elements $a_1, \cdots, a_N$.)

The probability distribution of R_h in this subpopulation can be determined as follows: Consider the set of $N!$ values of R_h which are obtained by substituting for $(x_1, \cdots, x_N)$ all possible permutations of $(a_1, \cdots, a_N)$. (A value which occurs more than once is counted as many times as it occurs.) Each of these values of R_h has the probability $1/N!$. On the basis of this distribution of R_h an exact test of significance can be carried out Suppose that α is the level of significance, i.e , the size of the critical region. We choose as critical region a subset of M values out of the set of $N!$ values of R_h where $M/N! = \alpha$. The sub-

[3] If the existence of a trend is feared it may be preferable to use the non-circular statistic discussed, for example, in [2].

set of M values which constitute the critical region will depend in each particular problem on the possible alternatives to randomness. For example, if a linear trend is the only possible alternative to randomness, then the critical region will consist of the M largest values[4] of R_h. The value of the lag h will also be chosen on the basis of the alternatives under consideration For instance, if some cyclical movement in the data is suspected the choice of h will depend on the form of these cycles. The general idea underlying the choice of the subset of M values and of the lag is to make the power of the test with respect to the alternatives which are particularly feared as high as possible.

If R_h has the same value for several permutations of $(a_1, \cdots, a_N)$, it may be impossible to have a critical region consisting of exactly M values of R_h. For example, if $a_1 = a_2 = \cdots = a_N$, then all the $N!$ values of R_h are equal, and the number of values of R_h included in the critical region must be either 0 or $N!$. If $F(x)$ is continuous the probability that two values of R_h be equal is zero. This explains why an exact test is always possible when $F(x)$ is continuous. On the other hand, if $F(x)$ is not continuous, the probability that several values of R_h be equal is positive However, the theorem we shall prove in Section 4 shows that in the limit an exact test is possible even when $F(x)$ is not continuous, but has finite moments and a positive variance. For if the latter is true, the probability is one that the weaker conditions for the validity of our theorem (given at the end of Section 4) will be fulfilled

Consider the statistic

$$\bar{R}_h = \sum_{\alpha=1}^{N} x_\alpha x_{h+\alpha} \tag{2}$$

where $x_{h+\alpha}$ is to be replaced by $x_{h+\alpha-N}$ for all values of α for which $h + \alpha > N$. Since in the subpopulation under consideration $\sum_{\alpha=1}^{N} x_\alpha$ and $\sum_{\alpha=1}^{N} x_\alpha^2$ are constants, the statistic $\bar{R}_h$ is a linear function of R_h in this subpopulation Hence, the test based on $\bar{R}_h$ is equivalent to the test based on R_h. Since $\bar{R}_h$ is simpler than R_h, in what follows we shall restrict ourselves to the statistic $\bar{R}_h$.

We shall now show that, if h is prime to N, the totality T_h of the $N!$ values taken by $\bar{R}_h$ is the same as T_1, the totality of the $N!$ values taken by $\bar{R}_1$.

In the argument which follows it is to be understood that, whenever a positive integer is greater than N, it is to be replaced by that positive integer less than or equal to N which differs from it by an integral multiple of N.

Clearly it will be sufficient to show the existence of a permutation $p_1, p_2, \cdots, p_N$ of the first N integers such that

$$p_i + 1 = p_{i+h} \qquad (i = 1, 2, \cdots, N).$$

Such a permutation is given by

$$j = p_{(j-1)h+1} \qquad (j = 1, 2, \cdots, N).$$

For if $j \neq j'$ then $(j-1)h + 1 \not\equiv (j'-1)h + 1$ because h is prime to N. Hence to every positive integer i there is a unique positive integer j, $(i, j \leq N)$ such

[4] See footnote 3

that

$$i = (j - 1)h + 1$$

Now

$$p_i + 1 = p_{(j-1)h+1} + 1 = j + 1 = p_{jh+1} = p_{i+h},$$

which is the required result.

In what follows we shall restrict ourselves to the case when h is prime to N. This is not a very restrictive assumption since in practice h will be small as compared with N and by omitting a few observations we can always make N prime to h. Since T_h is the same as T_1 we shall deal with the statistic $\bar{R}_1$ only. To simplify the notation we shall write R instead of $\bar{R}_1$. Thus, the test procedure will be based on the statistic

$$R = \sum_{\alpha=1}^{N-1} x_\alpha x_{\alpha+1} + x_N x_1 . \tag{4}$$

If N is very small an exact test of significance can be carried out by actually calculating the $N!$ possible values of R. However, this procedure is practically impossible if N is not small. In Section 3 the exact mean value and variance of R will be calculated, and in section 4 the normality of the limiting distribution of R will be proved. Thus, if N is sufficiently large so that the limiting distribution of R can be used, a test of significance can easily be carried out. Difficulties in carrying out the test arise if N is neither sufficiently small to make the computation of the $N!$ values of R practically possible, nor sufficiently large to permit the use of the limiting distribution. In such cases it may be helpful to determine the third and fourth, and perhaps higher, moments of R, on the basis of which upper and lower limits for the cumulative distribution of R can be derived. (For a description of the Tchebycheff inequalities by which this can be done see, for example, Uspensky, [8], pp. 373–380.) Since the limiting distribution is normal it may be useful to approximate the distribution by a Gram-Charlier series or to employ similar methods.

3. Mean value and variance of R.[5] It is clear that

$$\begin{aligned} E(R) &= NE(x_1 x_2) = \frac{N}{N(N-1)} \sum\sum_{\alpha \neq \beta} a_\alpha a_\beta \\ &= \frac{1}{N-1} [(a_1 + \cdots + a_N)^2 - (a_1^2 + \cdots + a_N^2)]. \end{aligned} \tag{5}$$

To calculate the variance of R we first calculate the second moment of R about the origin. We have

$$\begin{aligned} E(R^2) &= E(x_1 x_2 + \cdots + x_{N-1} x_N + x_N x_1)^2 \\ &= NEx_1^2 x_2^2 + 2NEx_1 x_2^2 x_3 + (N^2 - 3N) Ex_1 x_2 x_3 x_4 . \end{aligned} \tag{6}$$

[5] The first four moments of a similar statistic have been obtained by Young [3].

To express the expected values $Ex_1^2x_2^2$, $Ex_1x_2^2x_3$, and $Ex_1x_2x_3x_4$ we shall introduce the following notations for the symmetric functions of $a_1, \cdots, a_N$: For any set of positive integers $i_1, i_2, \cdots, i_k$ the symbol $S_{i_1 i_2 \cdots i_k}$ denotes the symmetric function $\sum_{\alpha_k} \cdots \sum_{\alpha_1} a_{\alpha_1}^{i_1} \cdots a_{\alpha_k}^{i_k}$ where the summation is to be taken over all possible sets of k positive integers $\alpha_1, \cdots, \alpha_k$ subject to the restriction that $\alpha_u \leq N$ and $\alpha_u \neq \alpha_v$ $(u, v = 1, \cdots, k)$.

From (6) we easily obtain

$$E(R^2) = \frac{N}{N(N-1)} S_{22} + \frac{2N}{N(N-1)(N-2)} S_{121} + \frac{N^2 - 3N}{N(N-1)(N-2)(N-3)} S_{1111} \tag{7}$$

$$= \frac{S_{22}}{(N-1)} + \frac{2S_{121}}{(N-1)(N-2)} + \frac{S_{1111}}{(N-1)(N-2)}.$$

It will probably facilitate computation to express each of the symmetric functions in the right member of (7) by a sum of terms, each a product of factors $S_r (r = 1, 2, \cdots)$. One can easily verify the relationships

$$S_{11} = S_1^2 - S_2 \tag{8}$$

$$S_{12} = S_{21} = S_1S_2 - S_3 \tag{9}$$

$$S_{13} = S_{31} = S_1S_3 - S_4 \tag{10}$$

$$S_{22} = S_2^2 - S_4 \tag{11}$$

$$S_{111} = S_{11}S_1 - 2S_{12} = (S_1^2 - S_2)S_1 - 2(S_1S_2 - S_3) \tag{12}$$

$$= S_1^3 - 3S_1S_2 + 2S_3$$

$$S_{112} = S_{121} = S_{211} = S_{11}S_2 - 2S_{13} \tag{13}$$

$$= (S_1^2 - S_2)S_2 - 2(S_1S_3 - S_4)$$

$$= S_1^2S_2 - S_2^2 - 2S_1S_3 + 2S_4$$

$$S_{1111} = S_{111}S_1 - 3S_{112} \tag{14}$$

$$= S_1^4 - 3S_1^2S_2 + 2S_1S_3 - 3S_1^2S_2 + 3S_2^2 + 6S_1S_3 - 6S_4$$

$$= S_1^4 - 6S_1^2S_2 + 8S_1S_3 + 3S_2^2 - 6S_4.$$

It follows from (5) that

$$E(R) = \frac{1}{N-1}(S_1^2 - S_2), \tag{15}$$

and from (7), (11), (13), (14), and (15) that the variance of R is given by

$$\sigma^2(R) = E(R^2) - [E(R)]^2$$

$$= \frac{S_2^2 - S_4}{N-1} + \frac{S_1^4 - 4S_1^2S_2 + 4S_1S_3 + S_2^2 - 2S_4}{(N-1)(N-2)} - \frac{1}{(N-1)^2}(S_1^2 - S_2)^2. \tag{16}$$

The mean value and variance of R can easily be computed from (15) and (16) as soon as the values of S_1, S_2, S_3, and S_4 have been determined.

The formulas (15) and (16) are considerably simplified if $S_1 = 0$. In the special case that $S_1 = 0$ we have

$$E(R) = -\frac{S_2}{N-1} \tag{15'}$$

and

$$\sigma^2(R) = \frac{S_2^2 - S_4}{N-1} + \frac{S_2^2 - 2S_4}{(N-1)(N-2)} - \frac{S_2^2}{(N-1)^2}. \tag{16'}$$

We can always make S_1 equal to zero by replacing a_α by $b_\alpha = a_\alpha - N^{-1}\Sigma a_\alpha$. This substitution is permissible, since it changes the statistic R only by an additive constant and consequently leaves the test procedure unaffected. Thus, in practical applications it may be convenient to replace a_α by b_α and to use formulas (15') and (16').

4. Limiting distribution of R. Let $\{a_\alpha\}$ ($\alpha = 1, 2, \cdots$ ad inf.) be a sequence of real numbers with the following properties:

a) There exists a sequence of numbers $A_1, A_2, \cdots, A_r, \cdots$ such that

$$\frac{1}{N}\left|\sum_{\alpha=1}^{N} a_\alpha^r\right| \leq A_r \qquad (r = 1, 2, \cdots \text{ ad inf.}) \tag{17}$$

for all N. (This condition means that the moments about the origin of the sequence $a_1, a_2, \cdots, a_N$ are bounded functions of N.)

b) If

$$\delta(N) = \frac{1}{N}\left[\sum_{\alpha=1}^{N} a_\alpha^2 - \frac{1}{N}\left(\sum_{\alpha=1}^{N} a_\alpha\right)^2\right],$$

then

$$\liminf_N \delta(N) > 0. \tag{18}$$

(This condition means that the dispersion of the N values $a_1, a_2, \cdots, a_N$ is eventually bounded below.)

Let $R(N)$ be the serial correlation coefficient R as defined in (4), where $x_1, \cdots, x_N$ is a random permutation of $a_1, a_2, \cdots, a_N$. We shall prove the following

THEOREM: *As $N \to \infty$, the probability that*

$$\frac{R(N) - E(R(N))}{\sigma(R(N))} < t$$

approaches the limit

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-\frac{1}{2}x^2}\,dx.$$

For any function $f(N)$ and any positive function $\phi(N)$ let

$$f(N) = O(\phi(N))$$

mean that $|f(N)/\phi(N)|$ is bounded from above for all N, and let

$$f(N) = \Omega(\phi(N))$$

mean that

$$f(N) = O(\phi(N))$$

and that $\liminf_N |f(N)/\phi(N)| > 0$. Also let

$$f(N) = o(\varphi(N))$$

mean that

$$\lim_{N\to\infty} \frac{f(N)}{\phi(N)} = 0.$$

Let $[\rho]$ denote the largest integer less than or equal to ρ.

To simplify the proof we shall temporarily assume:

c) There exists a positive constant K such that, for every positive integral N,

$$-K \leq S_1 = \sum_{\alpha=1}^{N} a_\alpha \leq K. \tag{19}$$

This restriction will be removed later.

LEMMA 1:

$$\sum_{\alpha_1<} \cdots \sum_{\cdots<\alpha_k} a_{\alpha_1} a_{\alpha_2} \cdots a_{\alpha_k} = O(N^{[\frac{1}{2}k]}).$$

PROOF: $\sum_{\alpha_1<} \cdots \sum_{\cdots<\alpha_k} a_{\alpha_1} \cdots a_{\alpha_k}$ can be written as the sum of a finite number of terms where each term is a product of factors S_r $(r = 1, 2, \cdots)$. This representation will be called the normal representation of $\sum \cdots \sum a_{\alpha_1} \cdots a_{\alpha_k}$. Since $S_1 = O(1)$ by (19) and $S_r = O(N)$ by (17) and since the number of factors S_r $(r > 1)$ in a single term of the normal representation of $\sum \cdots \sum a_{\alpha_1} \cdots a_{\alpha_k}$ is at most $[\frac{1}{2}k]$, the equation $\sum \cdots \sum a_{\alpha_1} \cdots a_{\alpha_k} = O(N^{[\frac{1}{2}k]})$ must hold.

LEMMA 2: *Let* $y = x_1 \cdots x_k z$, *where* $z = x_{k+1}^{i_1} \cdots x_{k+r}^{i_r}$ *and* $i_j > 1$ $(j = 1, \cdots, r)$. *If* $(x_1, \cdots, x_N)$ *is a random permutation of* $a_1, \cdots, a_N$, *and if* k, r, $i_1, \cdots, i_r$ *are fixed values independent of* N, *then* $E(y) = O(N^{[\frac{1}{2}k]-k})$.

PROOF: Let $E(y \mid x_{k+1}, \cdots, x_{k+r})$ be the conditional expected value of y when $x_{k+1}, \cdots, x_{k+r}$ are fixed. It follows easily from Lemma 1 that

$$E(y \mid x_{k+1}, \cdots, x_{k+r}) = O(N^{[\frac{1}{2}k]-k}).$$

Hence also $E(y) = O(N^{[\frac{1}{2}k]-k})$ and Lemma 2 is proved.

Denote $x_\alpha x_{\alpha+1}$ by $y_\alpha (\alpha = 1, \cdots, N-1)$ and $x_N x_1$ by y_N, and consider the expansion of $(y_1 + \cdots + y_N)^r$. Let y be a term of this expansion, i.e., $y = \frac{N!}{i_1! \cdots i_u!} y_{\alpha_1}^{i_1} \cdots y_{\alpha_u}^{i_u}$ $(\alpha_1 < \alpha_2 < \cdots < \alpha_u)$. We will say that two factors y_α and y_β are neighbors if $|\alpha - \beta + 1|$ or $|\alpha - \beta - 1|$ is either 0 or N. The set of u factors $y_{\alpha_1}, \cdots, y_{\alpha_u}$ can be subdivided into cycles as follows: The first cycle contains y_{α_1} and all those y_α which can be reached from y_{α_1} by a succession of neighboring y_α. The second cycle contains the first y_α of the remaining sequence and all those which can be reached from the first y_α by a succession of neighboring y_α. The third cycle is similarly constructed from the remaining sequence, etc. After a finite number of cycles have been withdrawn the sequence will be exhausted. If m is the number of such cycles we will say that y has m cycles.

LEMMA 3: *Let y be a term of the expansion $(x_1x_2 + \cdots + x_Nx_1)^r = (y_1 + \cdots + y_N)^r$ (r fixed). Let m be the number of cycles in y and k be the number of linear factors in y if y is written as a function of $x_1, \cdots, x_N$ (i.e., if we replace y_α by $x_\alpha x_{\alpha+1}$). Then the maximum value of $m + [\frac{1}{2}k] - k$ is equal to $[\frac{1}{2}r]$.*

PROOF: First we maximize $m + [\frac{1}{2}k] - k$ with respect to k when m is fixed. If $m \leq [\frac{1}{2}r]$, then the minimum value of k is obviously zero. Let $m = [\frac{1}{2}r] + r'$ $(r' > 0)$. The minimum value of k is reached if each cycle consists of a single factor y_α and if each factor y_α in y is either linear or squared. If r is even, then the minimum value of k is $4r'$ and if r is odd then the minimum value of k is $4r' - 2$. Hence for $m = [\frac{1}{2}r] + r'$ we have

$$\max_k (m + [\tfrac{1}{2}k] - k) = [\tfrac{1}{2}r] - r' \qquad \text{if } r \text{ is even}$$

and

$$= [\tfrac{1}{2}r] - r' + 1 \text{ if } r \text{ is odd.}$$

Hence maximizing with respect to m and k we obtain

$$\max (m + [\tfrac{1}{2}k] - k) = [\tfrac{1}{2}r],$$

and Lemma 3 is proved.

LEMMA 4: *The expected value of the sum of all those terms in the expansion of $(x_1x_2 + \cdots + x_Nx_1)^r$ for which m is the number of cycles and k the number of linear factors (if y is expressed in terms of $x_1, \cdots, x_N$) is equal to $O(N^{m+[\frac{1}{2}k]-k})$.*

This Lemma follows from Lemma 2 and the fact that the number of terms y with the required properties is $O(N^m)$.

LEMMA 5:

$$E(x_1x_2 + \cdots + x_Nx_1)^r = O(N^{[\frac{1}{2}r]}).$$

This follows from Lemmas 3 and 4.

LEMMA 6: *If r is even then*

$$E(x_1x_2 + \cdots + x_Nx_1)^r = \left(C_{\frac{1}{2}r}^N \left(\frac{r!}{2^{\frac{1}{2}r}}\right) E(x_1^2 x_2^2 \cdots x_r^2)\right) + o(N^{\frac{1}{2}r}).$$

PROOF: It follows easily from our considerations in proving Lemma 3 that $m + [\frac{1}{2}k] - k < \frac{1}{2}r$ for all terms in the expansion of $(x_1x_2 + \cdots + x_Nx_1)^r$ which are not of the type $x_1^2 \cdots x_r^2$. Hence it follows from Lemma 4 that the expected value of the sum of all those terms in the expansion of $[x_1x_2 + \cdots + x_Nx_1]^r$ which are not of the type $x_1^2 \cdots x_r^2$ is equal to $o(N^{\frac{1}{2}r})$. Lemma 6 follows from the fact that $2^{-\frac{1}{2}r}r!$ is the coefficient of the terms of the type $x_1^2 \cdots x_r^2$ in the expansion of $(x_1x_2 + \cdots + x_Nx_1)^r$ and that the number of terms of such type is equal to $C_{\frac{1}{2}r}^N$.

LEMMA 7. $\lim_{N\to\infty} \dfrac{E(x_1x_2 + \cdots + x_Nx_1)^r}{\{E(x_1x_2 + \cdots + x_Nx_1)^2\}^{\frac{1}{2}r}} = 0$ *if* r *is odd and* $= 2^{-\frac{1}{2}r}r!/(\frac{1}{2}r)!$ *if* r *is even.*

PROOF: From Lemma 6 it follows that

$$E(x_1x_2 + \cdots + x_Nx_1)^2 = NE(x_1^2x_2^2) + o(N) = \Omega(N). \tag{20}$$

The first half of Lemma 7 follows from Lemma 5 and equation (20). If r is even then it follows from (20) that

$$\begin{aligned} \lim \frac{E(x_1x_2 + \cdots + x_Nx_1)^r}{\{E(x_1x_2 + \cdots + x_Nx_1)^2\}^{\frac{1}{2}r}} &= \lim_{N\to\infty} \frac{2^{-\frac{1}{2}r}C_{\frac{1}{2}r}^N r!\, E(x_1^2 \cdots x_r^2)}{N^{\frac{1}{2}r}(Ex_1^2x_2^2)^{\frac{1}{2}r}} \\ &= \lim \frac{r!}{2^{\frac{1}{2}r}(\frac{1}{2}r)!} \frac{E(x_1^2 \cdots x_r^2)}{(E(x_1^2x_2^2))^{\frac{1}{2}r}}. \end{aligned} \tag{21}$$

It follows from (17), (19), and the normal representation of symmetric functions that

$$k! \sum_{a_{\alpha_1}<a_{\alpha_2}<\cdots<a_{\alpha_k}} \cdots \sum a_{\alpha_1}^2 \cdots a_{\alpha_k}^2 = S_2^k + O(N^{k-1}).$$

From (17) and (18) we have $S_2 = \Omega(N)$. Since

$$E(x_1^2 \cdots x_r^2) = r!\Big(\sum_{a_{\alpha_1}<a_{\alpha_2}<\cdots<a_{\alpha_r}} \cdots \sum a_{\alpha_1}^2 \cdots a_{\alpha_r}^2\Big)[N(N-1)\cdots(N-r+1)]^{-1},$$

we obtain

$$\lim_{N\to\infty} \frac{E(x_1^2 \cdots x_r^2)}{(E(x_1^2x_2^2))^{\frac{1}{2}r}} = 1. \tag{22}$$

The second half of Lemma 7 follows from (21) and (22).

LEMMA 8:

$$\lim_{N\to\infty} \frac{E(R(N))}{\sigma(R(N))} = 0, \tag{23}$$

$$\lim_{N\to\infty} \frac{E(R^2(N))}{\sigma^2(R(N))} = 1. \tag{24}$$

PROOF: Equation (24) is a trivial consequence of (23). From (15) $E(R) = O(1)$ and from (16) $\sigma(R) = \Omega(N^{\frac{1}{2}})$. The lemma follows easily from these relations.

PROOF OF THE THEOREM. According to Lemma 7 the r-th moment of $R[E(R^2)]^{-\frac{1}{2}}$ approaches the r-th moment of the normal distribution as $N \to \infty$. From this and Lemma 8 the required result follows if condition (c) holds. It remains therefore merely to remove condition (c) Assume now only that $a_1, a_2, \cdots, a_\alpha, \cdots$ satisfy conditions (a) and (b).

$R(N)$ is formed from the population of values $a_1, a_2, \cdots, a_N$. Addition of a constant q to $a_1, \cdots, a_N$ adds the same constant to all the values of $R(N)$ and hence leaves $[R(N) - E(R(N))]/\sigma(R(N))$ unaltered. Let $q^{(N)}$ be $-\sum_{\alpha=1}^{N} a_\alpha/N$ and write $b_\alpha^{(N)} = a_\alpha + q^{(N)}$. Consider the sequences

$$B^{(i)} = b_1^{(i)}, b_2^{(i)}, \cdots, b_i^{(i)} \qquad (i = 1, 2, \cdots, \text{ad inf.}).$$

From (17) it follows that the $| q^{(N)} |$ are bounded for all N. Hence the sequences $B^{(i)}$ satisfy condition (a). They obviously satisfy condition (c). Since $\delta(j)$ is invariant under addition of a constant we have

$$\liminf_{j} \frac{1}{j}\left(\sum_{\alpha=1}^{j} (b_\alpha^{(j)})^2 - \frac{1}{j}\left(\sum_{\alpha=1}^{j} b_\alpha^{(j)}\right)^2\right) > 0,$$

so that the $B^{(i)}$ satisfy condition (b). Since $[R(N) - E(R(N))]/\sigma(R(N))$ has the same distribution in the sequence $a_1, a_2, \cdots, a_N$ as in the sequence $B^{(N)}$, the theorem follows.

It should be remarked that the theorem remains valid if conditions (a) and (b) are replaced by the weaker condition

$$\mu_r/\mu_2^{\frac{1}{2}r} = O(1) \qquad (r = 3, 4, \cdots, \text{ad inf.})$$

where

$$\mu_r = \frac{1}{N}\sum_{\alpha=1}^{N}\left(a_\alpha - \frac{1}{N}\sum_{\alpha=1}^{N} a_\alpha\right)^r.$$

This follows easily from the fact that $[R(N) - E(R(N))]/\sigma(R(N))$ remains unaltered if we replace the sequence $a_1, \cdots, a_N$ by the sequence $c_1^N, c_2^N, \cdots, c_N^N$ where

$$c_\alpha^N = \left(a_\alpha - \frac{1}{N}\sum_{1}^{N} a_\alpha\right) \Big/ \left[\frac{1}{N}\sum\left(a_\alpha - \frac{1}{N}\sum a_\alpha\right)^2\right]^{\frac{1}{2}}.$$

Conditions (a) and (b) are obviously satisfied by the sequence $c_1^N, \cdots, c_N^N$.

5. Transformation of the original observations.

Let $f(t)$ be a continuous and strictly monotonic function of t $(-\infty < t < +\infty)$. Suppose we replace the original observations $a_1, \cdots, a_N$ by $d_1, \cdots, d_N$, where $d_\alpha = f(a_\alpha)$ $(\alpha = 1, \cdots, N)$. We obtain a valid test of significance if we carry out the test procedure as if $d_1, \cdots, d_N$ were the observed values instead of $a_1, \cdots, a_N$. We could also replace the observed values $a_1, \cdots, a_N$ by their ranks. The question arises whether there is any advantage in making the test on the transformed values instead of on the original observations. It may well

be that by certain transformations we could considerably increase the power of the test with respect to alternatives under consideration. This problem needs further study.

6. Summary. A test procedure based on serial correlation is given for testing the hypothesis that $x_1, \cdots, x_N$ are independent observations from the same population, i.e., that $x_1, \cdots, x_N$ is a random series. By considering the distribution of the serial correlation coefficient in the subpopulation consisting of all permutations of the actually observed values a test procedure is obtained such that

a) if the common c.d.f. $F(x)$ is continuous, the size of the critical region, i.e., the probability of rejecting the hypothesis of randomness when it is true, does not depend upon $F(x)$,

b) if $F(x)$ is not continuous but all its moments are finite and its variance is positive, the size of the critical region approaches, as $N \to \infty$, the value it would have if $F(x)$ were continuous. Thus in the limit an exact test is possible in this case as well.

It is shown that the test based on the serial correlation with lag h is equivalent to the test based on the statistic[6]

$$\sum_{\alpha=1}^{N} x_\alpha x_{h+\alpha}$$

where $x_{h+\alpha}$ is to be replaced by $x_{h+\alpha-N}$ for all values of α for which $h + \alpha > N$. If h is prime to N, the distribution of $\sum_1^N x_\alpha x_{h+\alpha}$ is exactly the same as the distribution of $R = \sum_1^N x_\alpha x_{1+\alpha}$.

The mean value and variance of R are given by the following expressions:

$$E(R) = (S_1^2 - S_2)/(N - 1)$$

and

$$\sigma^2(R) = \frac{S_2^2 - S_4}{N-1} + \frac{S_1^4 - 4S_1^2 S_2 + 4S_1 S_3 + S_2^2 - 2S_4}{(N-1)(N-2)} - \frac{(S_1^2 - S_2)^2}{(N-1)^2}$$

where $S_r = x_1^r + \cdots + x_N^r$.

It is shown that under some mild restrictions the limiting distribution of R is normal. The test procedure can therefore be easily carried out when N is sufficiently large to permit the use of the limiting distribution of R.

REFERENCES

[1] R. L. Anderson, *Annals of Math. Stat.*, Vol. 13 (1942), p. 1.
[2] T. Koopmans, *ibid.*, p. 14.
[3] L. C. Young, *Annals of Math. Stat.*, Vol. 12 (1941), p. 293.
[4] J. v. Neumann, *Annals of Math. Stat.*, Vol. 12 (1941), p. 367.
[5] J. v. Neumann, *Annals of Math. Stat.*, Vol. 13 (1942), p. 86.
[6] B. I. Hart and J. v. Neumann, *Annals of Math. Stat*, Vol. 13 (1942), p. 207.
[7] J. D. Williams, *Annals of Math. Stat.*, Vol 12 (1943), p. 239.
[8] J. Uspensky, *Introduction to Mathematical Probability*, New York, 1937.

[6] If the non-circular definition of the serial correlation coefficient is used, the term $x_N x_{N+h}$ should be omitted.

ON A GENERAL CLASS OF "CONTAGIOUS" DISTRIBUTIONS

By W. Feller

Brown University

1. Introduction. In a paper of considerable interest, J. Neyman [11] recently discussed frequently occurring situations where the usual tests of significance fail. He discussed, in particular, experiences in entomology and bacteriology which cannot be described by the usual distribution functions and he constructed several new types of apparently contagious distributions. Now at first glance Neyman's investigation may seem of a rather specialized nature, and his distributions of a restricted applicability. It may therefore be useful to point out that they are intimately related to results obtained by various authors in connection with topics having so little apparent relation as accident statistics, telephone traffic, fire damage, sickness- and life-insurance, risk theory, and even an engineering problem. Viewed in the proper light of a general theory, Neyman's method is particularly closely related to some too little known considerations by Greenwood and Yule [6]. These authors were the first to find, and apply, the distribution which shortly afterwards was independently rediscovered by Eggenberger and Polya[1] [3, 4].

Greenwood and Yule discussed two types of what may conveniently be called contagion: with one type there is *true contagion* in the sense of Polya and Eggenberger, where each "favorable" event increases (or decreases) the probability of future favorable events; with the second type the events are, strictly speaking, independent and an *apparent contagion* is actually due to an inhomogeneity of the population. The two explanations are very different in nature as well as in practical implications. It is therefore most remarkable that Greenwood and Yule found their distribution assuming an *apparent contagion*; in their opinion this distribution *contradicts* true contagion. On the contrary, Polya and Eggenberger arrived at the *same* distribution assuming *true* contagion, while the possibility of an apparent contagion due to inhomogeneity seems not to have been noticed by them. The Greenwood-Yule-Polya-Eggenberger distribution has found many applications.[2] Therefore the possibility of its interpretation in two ways, diametrically opposite in their nature as well as in their implications is of greatest statistical significance. This fact is, incidentally, a justification for general theories in statistics.

We shall see that Neyman's contagious distributions belong to the *second* type and are related to the Polya-Eggenberger distribution only if the latter is

[1] The fact that the Polya-Eggenberger distribution is identical with the Greenwood-Yule distribution seems to be mentioned in the literature only in a Stockholm thesis by O. Lundberg [9].

[2] Of quite recent applications we mention Kitagawa and Huruya [8], Rosenblatt [15], O. Lundberg [9]. Only the latter seems aware of the double nature of the distribution.

interpreted in the sense of Greenwood and Yule. In Neyman's case as well as in the other cases referred to above we are concerned with inhomogeneous populations and there exists an extremely simple device to describe such situations appropriately. Once stated, this device will appear trivial. Nevertheless, a straightforward application of it would have avoided considerable mathematical difficulties in the literature and, occasionally, yielded better and simpler results. It seems also the simplest description of the mechanism behind many observed distributions, and therefore suited for a theory of tests[3].

To start in a purely formal manner, consider an arbitrary cumulative distribution function (c.d.f.) $F(x, a)$, depending on a parameter a, and another c.d.f. $U(a)$. *Then*

$$G(x) = \int F(x, a)\, dU(a) \tag{1.1}$$

(the integration extending over the domain of variation of a) *is again a c.d.f.* If, in particular, $U(a)$ is a step function, (1.1) reduces to

$$G(x) = \Sigma p_i F(x, a_i), \tag{1.2}$$

where p_i is the weight attached to a_i (we have, of course, $p_i \geq 0$, $\Sigma p_i = 1$). Instead of (1.2) one can write more simply

$$G(x) = \Sigma p_i F_i(x), \tag{1.3}$$

where the $F_i(x)$ are arbitrary c.d.f.'s. Of course, $F(x, a)$ and $U(a)$ may depend on additional parameters, and the procedure can be repeated.

The statistical meaning of (1.3) is clear. Consider a population made up of several subgroups $A_1, A_2, \cdots$, mixed at random in proportions $p_1 : p_2 : \cdots$. If $F_i(x)$ is the c.d.f. of some character in A_i, then $G(x)$, as defined by (1.3), will represent the c.d.f. of that character in the total population, provided that the subgroups A_i are statistically independent. Similarly (1.1) describes an infinitely composite population. Postponing a discussion of the property of contagion to the last section, we shall first deduce a few properties of the *compound Poisson-distribution*, considered first by Greenwood and Yule. Neyman's "Contagious Distributions of Type A" as well as the Polya-Eggenberger distribution belong to this class. Our next example of a special case of (1.1) is what F. E. Satterthwaite [16] called the "Generalized Poisson Distribution." It has been independently discovered by many authors and represents heterogeneity of quite different a nature. Instead of further examples we shall, in the fourth section, show how Neyman's most general contagious distribution can be deduced by a repeated application of (1.1).

[3] Incidentally, attention may be drawn to an argument by Greenwood and Yule showing that the χ^2-test when applied to the Poisson distribution is biased and tends to exaggerate the goodness of fit. The argument could be amplified from other experience.

Notation: If $F(x)$ and $G(x)$ are the c d.f.'s of two independent variates X and Y, then their *convolution*, (that is to say the c.d.f. of $X + Y$) will be denoted by $F(x)*G(x)$. Thus

$$F(x)*G(x) = \int_{-\infty}^{+\infty} F(x - y)\, dG(y). \tag{1.4}$$

More particularly we shall write

$$F(x)*F(x) = F^{2*}(x), \tag{1.5}$$
$$F^{n*}(x)*F(x) = F^{(n+1)*}(x).$$

We shall denote by $E(x)$ the *unitary c.d.f.*

$$E(x) = \begin{cases} 0 & \text{for } x < 1, \\ 1 & \text{for } x \geq 1, \end{cases} \tag{1.6}$$

so that $E^{n*}(x) = 0$ for $x < n$, and 1 for $x \geq n$.

2. The compound Poisson distribution. Consider the well-known Poisson expression

$$\pi(n;\, a) = e^{-a}\frac{a^n}{n!}, \tag{2.1}$$

where the parameter $a > 0$ gives the expected number of "events". We shall refer to (2.1) as the *simple Poisson distribution*. If different individuals of a population are associated with different values of a, and if the character a is distributed according to the cumulative probability law $U(a)$, the probability of n events in the total population will be given by

$$\pi_n = \int_0^{\infty} e^{-a}\frac{a^n}{n!}\, dU(a). \tag{2.2}$$

Following Greenwood and Yule we shall refer to (2 2) as the *compound Poisson distribution*. Referring for an interpretation to the last section, we first consider a few special cases.

a) If $U(a)$ is a step function we are led to expressions of the form

$$\pi_n = \frac{1}{n!}\sum_i p_i e^{-a_i} a_i^n. \tag{2.3}$$

Such a distribution has been successfully applied by C. Palm [12] to problems of telephone traffic, and by O. Lundberg [9] to sickness statistics.

b) If $U(a)$ is a Pearson Type III distribution

$$U'(a) = \left(\frac{1}{d}\right)^{\frac{h}{d}} \frac{1}{\Gamma\left(\frac{h}{d}\right)}\, e^{\frac{-a}{d}} x^{\frac{a}{d}-1} \tag{2.4}$$

(with $d > 0, h > 0$), then

$$\pi_n = \frac{1}{n!} \frac{\Gamma\left(n + \frac{h}{d}\right)}{\Gamma\left(\frac{h}{d}\right)} (1 + d)^{\frac{-d}{h}} \left(\frac{1}{1 + d}\right)^n. \tag{2.5}$$

This is the *Polya-Eggenberger distribution* in its usual form, and has in this form (with a slight change of notations) been derived by Greenwood and Yule.

c) If a takes on the values kc only, where $c > 0$ is a constant and $k = 0, 1, \cdots$, and if a is distributed according to the Poisson law

$$Prob\ \{a = kc\} = e^{-\lambda} \frac{\lambda^k}{k!}, \tag{2.6}$$

then

$$\pi_n = e^{-\lambda} \frac{c^n}{n!} \sum_{k=0}^{\infty} \frac{k^n}{k!} (e^{-c}\lambda)^k. \tag{2.7}$$

This is Neyman's *contagious distribution of type A* depending on two parameters (cf. section 4). If, instead, a is distributed according to a multiple Poisson law of form (2.3) we arrive at Neyman's *more-parametric distribution of type A*. They are, of course, essentially linear combinations of expressions of form (2.7).

It follows from the theory of Laplace transforms that *two compound Poisson distributions associated with different c.d.f.'s $U(a)$ are never identical.*

The compound Poisson distribution gives a simple explanation of a phenomenon recorded by Neyman and observable in many instances. In the experiments described by Neyman "the attempts to fit the Poisson Law $\cdots$ failed almost invariably with the characteristic feature that, as compared with the Poisson Law, there were too many empty plots and too few plots with only one larva". It is easily checked in the literature that similar situations arise frequently. Now the Poisson distribution is usually fitted by the method of moments. Accordingly, the compound Poisson law (2.2) ought to be compared with the simple Poisson distribution with the same mean value. The mean value of (2.2) is

$$m = \int_0^{\infty} a\, dU(a), \tag{2.8}$$

so that (2.2) ought to be compared with the Poisson distribution $\pi(n; m)$. Now, *whatever the c.d.f. $U(a)$, we have always*

$$\pi_0 \geq \pi(0, m) \tag{2.9}$$

and

$$\frac{\pi_1}{\pi_0} \leq m = \frac{\pi(1, m)}{\pi(0, m)}. \tag{2.10}$$

As a matter of fact, using Lagrange's form for the remainder in Taylor's formula, we have

$$\begin{aligned} \pi_0 &= e^{-m} \int_0^\infty e^{m-a}\, dU(a) \\ &\geq e^{-m} \int_0^\infty \{1 + (m - a)\}\, dU(a) = e^{-m} = \pi(0, m), \end{aligned} \tag{2.11}$$

which proves (2.9). Similarly

$$\begin{aligned} m\pi_0 - \pi_1 &= e^{-m} \int_0^\infty e^{m-a}(m - a)\, dU(a) \\ &\geq e^{-m} \int_0^\infty (m - a)\, dU(a) = 0, \end{aligned} \tag{2.12}$$

which proves (2.10).

The above theorem shows that, *whenever the material under observation is not quite homogeneous so that the compound Poisson law applies instead of the simple one, there will be too many cases with "no event" and, as compared with these cases, too few with "one event"*. It should be noticed, however, that it is *not* strictly true that *always*

$$\pi_1 < \pi(1, m). \tag{2.13}$$

As a matter of fact, even in the numerical example given by Neyman, the computed value π_1 exactly equals the observed value. Still, the inequality (2.13) will hold whenever the third moment about the mean of $U(a)$ is smaller than twice the second. Writing

$$\begin{aligned} \sigma^2 &= \int_0^\infty (a - m)^2\, dU(a), \\ M &= \int_0^\infty (a - m)^3\, dU(a), \end{aligned} \tag{2.14}$$

and using two more terms in the Taylor development of e^{m-a} than in (2.11) and (2.12) we see that

$$\pi_0 \geq e^{-m}\left\{1 + \frac{\sigma^2}{2} - \frac{1}{6}M\right\} \tag{2.15}$$

and

$$m\pi_0 - \pi_1 \geq e^{-m}\{\sigma^2 - \tfrac{1}{2}M\}. \tag{2.16}$$

These inequalities are slightly sharper than (2.9) and (2.10), and often permit us to estimate the variance of $U(a)$.

We note furthermore that the *variance of the compound Poisson distribution is*

$$\sigma^2 + m \tag{2.17}$$

as compared with the variance m of the corresponding simple Poisson distribution. Finally the following important property of the compound distribution may be mentioned: *Consider two independent variates X and Y distributed according to two compound Poisson distributions* $\{\pi_n^{(1)}\}$ *and* $\{\pi_n^{(2)}\}$ *associated with the c.d.f.'s* $U_1(a)$ *and* $U_2(a)$, *respectively. Then the variate* $X + Y$ *is distributed according to a compound Poisson law* $\{\pi_n\}$ *associated with the c.d.f.* $U(a) = U_1(a) * U_2(a)$ (cf. (1.4)).

It suffices to note that $U_i(a) \equiv 0$ for $a < 0$, so that

$$U(a) = \int_0^a U_1(a - s)\, dU_2(s);$$

therefore, after a permitted change of the order of integration

$$\begin{aligned}
\pi_n &= \int_0^\infty e^{-a} \frac{a^n}{n!} dU(a) \\
&= \int_0^\infty dU_2(s) \int_s^\infty e^{-a} \frac{a^n}{n!} dU_1(a - s) \\
&= \int_0^\infty dU_2(s) \int_0^\infty e^{-(s+t)} \frac{(s+t)^n}{n!} dU_1(t) \\
&= \sum_{k=0}^{n} \frac{1}{k!} \frac{1}{(n-k)!} \pi_k^{(1)} \pi_{n-k}^{(2)};
\end{aligned}$$

the last expression represents the convolution of $\{\pi_n^{(1)}\}$ and $\{\pi_n^{(2)}\}$.

Neyman's distributions of type A with two parameters are special cases of a compound Poisson process where $U(a)$ is a step function with jumps at equidistant places, the jumps being given by a simple Poisson distribution $\{\pi(n; \lambda)\}$. Now the convolution of two such distributions is again a simple Poisson distribution $\{\pi(n; 2\lambda)\}$ with jumps at the same places; hence the convolution of two distributions of type A is again a similar distribution with one parameter doubled.

As mentioned before, the notion of a compound Poisson distribution is due to Greenwood and Yule [6]. The time dependent compound Poisson process has been the object of detailed investigations by J. Dubourdieu [2] and O. Lundberg [9]. The latter has discussed also the problem of fitting the compound Poisson process to empirical distributions.

3. The generalized Poisson distribution. Let $F(x)$ be an arbitrary c.d.f. Then its n-fold convolution $F^{n*}(x)$ (cf. (1.5)) may be considered as a c.d.f. depending on a parameter n. Choosing, for the latter, the simple Poisson distribution (2.1) and performing the operation indicated in (1.1), we arrive at the c.d.f. of *the generalized Poisson law*

$$G(x) = \sum_{n=0}^{\infty} e^{-a} \frac{a^n}{n!} F^{n*}(x). \tag{3.1}$$

If, in particular, $F(x)$ is the unitary function (1.6), we have the ordinary Poisson law

$$\Pi(x) = \sum_{n=0}^{\infty} e^{-a} \frac{a^n}{n!} E^{n*}(x) = \sum_{n=0}^{[x]} e^{-a} \frac{a^n}{n!} \tag{3.2}$$

in its cumulative form.

The most frequently encountered application of the generalized Poisson distribution is to problems of the following type. Consider independent random events for which the simple Poisson distribution may be assumed, such as. telephone calls, the occurrence of claims in an insurance company, fire accidents, sickness, and the like. With each event there may be associated a random variable X. Thus, in the above examples, X may represent the length of the ensuing conversation, the sum under risk, the damage, the cost (or length) of hospitalization, respectively. To mention an interesting example of a different type, A. Einstein Jr. [5] and G. Polya [13, 14] have studied a problem arising out of engineering practice connected with the building of dams, where the events consist of the motions of a stone at the bottom of a river; the variable X is the distance through which the stone moves down the river.

Now, if $F(x)$ is the c d.f of the variable X associated with a single event, *then* $F^{n*}(x)$ is the c.d.f. of the accumulated variable associated with n events. Hence (3.1) is the probability law of the sum of the variables (sum of the conversation times, total sum paid by the company, total damage, total distance travelled by the stone, etc.).

In view of the above examples, it is not surprising that the law (3.1), or special cases of it, have been discovered, by various means and sometimes under disguised forms, by many authors. Quite recently Satterthwaite [16] was led to it (in the above simple form) from problems in insurance. Related (but less elegant) considerations may be found in a paper by W. G. Ackermann [1]. Simple as they are, the above considerations leading to (3.1) furnish a complete solution of the problem in all the cases mentioned. Unfortunately, the special features of the problems often so overshadow the essential point, that one is often led to unnecessarily complicated and incomplete solutions. As an example of the difficulties in considering special cases we mention that Polya [13, 14] was led to a partial differential equation of the hyperbolic type, which conceals the elementary nature of the problem.

If $F(x)$ is itself a Poisson c.d.f. (3.1) reduces to (2.7) *Thus Neyman's distribution of type A depending on two parameters is both a compound and a generalized Poisson distribution.* We shall later on see that the generalized Poisson distribution plays an even more important rôle in Neyman's theory

The main properties of (3.1) are easily derived using characteristic functions. If $\varphi(z)$ is the characteristic function of $F(x)$, the characteristic function of $G(x)$ is

$$\psi(z) = e^{a(\varphi(z)-1)}. \tag{3.3}$$